

## Conscience call

**Nuclear proliferation remains a potent threat — and scientists' active engagement is essential if it is to be effectively addressed.**

**N**uclear weapons are little more than a footnote in most physics textbooks. But they live on and seem to be spreading. As this week's issue details (see pages 432 and 441), a decades-old treaty designed to stop the proliferation of nuclear weapons is beginning to unravel. All around the world, scientists are gaining the knowledge and technology they need to build a bomb. They may not be working to build a nuclear weapon today, but they are ready, should their governments ever call on them to do so in the future.

Politicians cannot be depended upon to solve this growing problem on their own. They will tend to decry the nuclear advances of their adversaries, while turning a blind eye to the progress of their friends. So it was last week, when US Secretary of State Colin Powell openly denounced the nuclear ambitions of Iran while passing through Brazil, a US ally whose own more advanced nuclear programme he has openly supported.

More than politicians, scientists have always understood the destructive power of the bomb as something distinct from its owner. The physicists who gave birth to the device recognized early on that it had the potential to ruin humanity, and many of them, including Albert Einstein and Robert Oppenheimer, took powerful, public positions against nuclear weapons after the end of the Second World War.

Like those distinguished scientists, modern researchers must take up the important work of disarmament and have much to offer in doing so. They have technical expertise that can be brought to bear on several critical issues related to stopping proliferation. In addition to developing the detectors that many governments seek to protect their ports and cities, researchers can think of techniques that further disarmament. For example, if an international treaty regulating fissile

materials is ever to be created, as many arms-control experts hope, then credible technologies to track and inventory such materials must be developed.

Moreover, scientists have a unique opportunity to inform the public about the risks of nuclear weapons. Since the end of the cold war, public awareness of the world's massive nuclear arsenals has faded, even though thousands of weapons remain on hair-trigger alert. It is up to physicists, chemists, computer scientists, and all those who understand what goes into nuclear weapons, to remind the public of the devastating power of the bomb and the constant threat it poses.

Finally, and most importantly, scientists must bridge the nationalist and cultural barriers that stand in the way of disarmament. In the cold war, the scientific exchange and mutual respect shared by US and Soviet researchers helped pave the way for treaties to reduce the number of nuclear weapons held by both nations. Today, a different gap exists between developed nuclear states, such as the United States, France and Britain, and developing nations with nuclear weapons or ambitions, such as India, Pakistan and Iran. Once again, it is the scientists who must open up lines of communication and begin work towards disarmament. Together, they need to help the world arrive at a new global consensus: that nuclear weapons — regardless of whose hands they are in — pose a threat to every nation's security.

Is it realistic to expect that the world can someday be free of the bomb? Perhaps not, but researchers everywhere owe it to future generations to move us further in that direction. In the words of the great Soviet weaponeer-turned-humanitarian, Andrei Sakharov: "The conscience of the contemporary scientist cannot distinguish the suffering of his contemporaries and that of posterity." ■

## Google Nouveau

**On the Internet, 2004 promises to be a vintage year for searching.**

**O**n the third Thursday of November, wine enthusiasts the world over traditionally race to be first to try out *l'arrivée* of *Beaujolais Nouveau*, a barely fermented wine that has rewarded harvest workers for seven centuries. This year, the same Thursday saw a second race as scientists and other information addicts rushed to try out a new vintage: 'Google Scholar', a test version of Google search that restricts web searches to academic texts (see page 423).

That the arrival of Google Scholar is an 'event' can be sensed from the waves of giddy excitement across Internet discussion groups as people posted their own musings on the free service. By its sheer scale and functionality, Google Scholar is a disruptive technology heralding profound changes in the way that many will access scientific information.

Google deserves congratulations. With a speed that only the private sector can deliver, it has addressed the most basic need of users accessing the literature: the ability to search the full text of a broad swathe of scientific information, cutting across publishers, journals and other sources. Two reservations: some publishers have not made their text available, and Google has not yet indexed all of the text

made available to it by others, including Nature Publishing Group.

But Google's database is orders of magnitude bigger than anything that exists today, and its search algorithms comparatively of lightning speed. The top search hits are the most cited — your search results won't be clogged by pages of poorly cited papers — and, in a further tweak, the hits at the top are those that are most cited by journals that themselves are the most cited.

Some worriers, including more than a few librarians, fear that Google Scholar will push scientists and students further towards what one describes as a "trend towards sloppy Internet-based research", instead of professional searches using abstracting and indexing databases, often bought at great cost by libraries. There is a grain of legitimate, if exaggerated, concern here. But one lesson of the Internet is that expensive, centralized ways of doing things tend to get undercut.

Google Scholar is far from perfect — it lacks basic functions such as search by date, characteristic of more costly databases. But the latest offering from *Château* Google is young, and will no doubt spur further innovation and competition in search. We at *Nature* drink to that, and look forward to the surprises that future years will bring. ■





**Free standing**  
NIH director  
defends plans for  
open-access archive  
p424



**Game plan**  
Billion-dollar study  
intends to track  
100,000 children  
p425



**Tissue issue**  
Biobanks urged  
to standardize  
sample accounts  
p426



**Fired up**  
Autistic arsonist  
convicted over  
pollution protest  
p428

# Science searches shift up a gear as Google starts Scholar engine

**Declan Butler**

Google has unveiled a test version of a search engine aimed specifically at academic material. First impressions of Google Scholar from librarians and computer experts suggest that the service is impressive in both scale and functionality.

"Google has made a very good start," says Ed Pentz, executive director of CrossRef, a collaboration of more than 1,000 publishers and societies that aims to improve online linking and access to papers.

The engine searches only research publications such as journal articles, books, preprints and technical reports, putting the most pertinent articles at the top of its searches by means of algorithms similar to those used by the firm's conventional web search. These analyse the number and importance of links pointing to sites (see *Nature* **405**, 112–115; 2000). In Google Scholar, papers with many citations are generally ranked highest, and they get a further boost if they are referenced by highly cited articles.

A test drive of the beta engine, or test version, provides a peek under the hood. A conventional Google search for "human genome" throws back several million hits, with genome centres and databases ranked top. Google Scholar, by contrast, returns only around 100,000, with the landmark *Science* and *Nature* human-genome papers from 2001 both appearing in the top three.

Google says almost all "major publishers" have allowed the full text of their papers to be searched, although it declined to provide a list of those involved. The engine also searches abstracts from online archives such as PubMed and the NASA Astrophysics Data System, and the complete text of physics preprints on the arXiv server. In total, almost half a billion documents are thought to be covered.

The index is itself still highly incomplete, however. For technical reasons, large swathes of the often complex article databases supplied to Google by collaborating publishers



Lord of the files: Anurag Acharya created Google Scholar to search academic papers.

are not covered. Google says it is working to solve these problems.

In addition, many papers can be only partially searched by the engine. Elsevier, the largest scientific publisher, has so far declined to allow Google to index its text, although the engine includes hits for more than a million Elsevier articles indexed as abstracts. "Google Scholar is an experiment, a beta version, and we are curious to see how it will be utilized and developed," says Marike Westra, a spokeswoman for the publisher.

## Scope for competition

Elsevier is also marketing a commercial search engine, known as Scopus (see *Nature* **428**, 683; 2004); annual institutional subscriptions to this engine run from \$25,000 to several hundred thousand dollars. The company argues that research institutions are willing to pay for high-quality search and information services such as Scopus and Web of Science, which is marketed by Thomson ISI of Philadelphia.

Some publishers may be concerned that Google Scholar has a subversive feature. Clicking on a hit returned by the engine takes the user to the article on the publisher's site. But Google Scholar also links to free versions

of the article archived on other sites, such as authors' personal home pages. It is unclear how publishers will respond to this.

The creative force behind Google Scholar is Anurag Acharya, an Indian-born computer scientist who was on the faculty at the University of California, Santa Barbara, before he joined the company in 2000. Acharya first had to make his software identify and gather scientific papers from around the web using simple rules based on the common format of scientific papers, and then extract the title, abstract, authors and references.

Extracting references, which come in a variety of formats and are often full of mistakes, is key. Once references and papers are interlinked, it is relatively simple to apply algorithms to create indexes and rankings.

"I started the project because I wanted to build something better for researchers," says Acharya. Building automated citation indexes was new to him, but the scaling up was helped by his background in designing large-scale distributed computing systems. "Extracting information and references was the hard part," he says. "Building an index, making it run fast, and stable, that was easy; I already know how to do all that."

► <http://scholar.google.com>

GOOGLE

# Lean budget leaves scientists wanting more

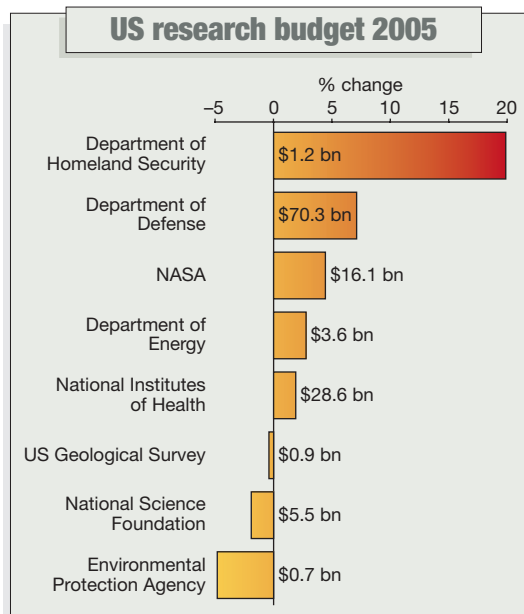
**Geoff Brumfiel and  
Tony Reichhardt, Washington**

US science agencies will face a financial squeeze in 2005. The federal budget, passed by Congress on 20 November, provides little new money for research and imposes cuts on several agencies.

The Department of Homeland Security and the Pentagon are the only clear winners (see chart) — both receive a big increase to develop technologies to combat terrorism. The National Institutes of Health gets a mere 2% increase to \$28.6 billion, well below the 15% increases it received annually between 1998 and 2003. Faringslightly better is the Department of Energy's science office, which sees a 2.8% increase to \$3.6 billion. On the losing side is the Environmental Protection Agency, whose research budget falls 4.8% to \$744 million.

But the National Science Foundation (NSF) was by far the most disappointed — its budget will drop 1.9% to \$5.5 billion next year. The decline, the agency's first since 1990, comes after several years of strong increases that science lobbyists had hoped would lead to a doubling of the foundation's budget. "I think various science coalitions are trying to scale down their expectations," says Nadine Lymn, director of public affairs at the Ecological Society of America in Washington.

The NSF budget cuts result in part from



President George W. Bush's plan to send astronauts to the Moon and eventually to Mars, says Michael Lubell, director of public affairs for the American Physical Society in Washington. That plan moved closer to reality last week, as NASA was granted its full request of \$16.1 billion, up 4.5% from last year. The funding should allow NASA to stay on schedule and pick contractors in August to build the first major piece of hardware for this

programme: a manned space vehicle that is expected to make a test flight in 2008.

Lubell sharply criticizes the decision to fund what he describes as a politically initiated project at the expense of peer-reviewed programmes at the NSF. And on 22 November, the American Physical Society issued a report warning that important projects could be delayed or cancelled as NASA prioritizes the Moon-Mars mission. "No one has laid out what the scientific benefits of this mission are going to be," says Lubell.

Funding for a few specific science programmes is also significantly reduced. Most notably, a White House miscalculation led Congress to assign just \$577 million to the construction of a nuclear-waste repository in Yucca Mountain, Nevada, well below the \$880 million requested for the project for 2005 (see *Nature* 430, 820; 2004).

Kei Koizumi, who directs the budget and policy programme at the American Association for the Advancement of Science in Washington, says tight years lie ahead for federal science agencies. The Bush administration is committed to halving the budget deficit by 2009, but it is still running up debt owing to recent tax cuts and military action in Iraq. "There's a lot of spending going on," Koizumi says, so the money must come from cuts to domestic programmes, such as science. ■

## NIH head stands firm over plans for open access

**Meredith Wadman, Washington**

The director of the US National Institutes of Health (NIH) has hit back at critics of his proposal for a freely accessible literature archive.

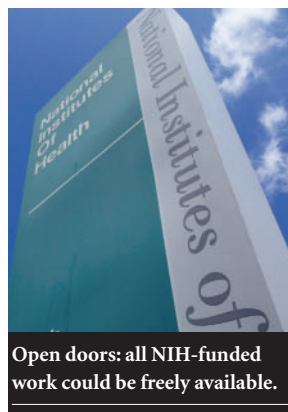
In an interview with *Nature*, Elias Zerhouni accused scientific publishers of floating "doomsday scenarios" in which the archive causes widespread cancellations of journal subscriptions and drives smaller publications out of business. He stressed that submission of papers on NIH-funded research would be left to the discretion of authors. But publishers rejected his assurances, saying that researchers would feel pressured to submit their papers for fear of losing out on future NIH grants.

Zerhouni's comments came on 16 November, at the end of a 60-day public discussion period that generated more than 6,000 comments on the proposal (see *Nature* 431, 115; 2004). The policy calls for all papers produced with NIH funding to be submitted electronically to the agency after completing peer review. Six months after publication,

the papers would appear in PubMed Central, the NIH's online public archive.

Many scientific publishers oppose the proposal, saying that offering their articles for free could drive journals out of business. They add that inaccuracies will be preserved in PubMed, because the policy proposes that articles be posted before copyediting and correcting. But patient-activist groups and librarians have been vocally supportive. They argue that the archive would improve public education, communication between scientists and the translation of biomedical advances into healthcare.

Zerhouni last week dismissed some of the publishers' fears, accusing them of releasing "misinformation" about the impact on subscriptions. He added that researchers would be free to opt out of the archive. "I'm



willing to take the risk of seeing the decision made not by government fiat but by the scientists themselves," he said. "If they don't wish to publish on the NIH website, that's their decision and the decision of their publishers, not mine."

Publishers say that the argument is disingenuous. They point out that the policy requires authors and not journals to submit papers. "Researchers would

be concerned that if they did not comply with this plan, they might be looked upon with less favour for awards of future NIH grants," says Allan Adler, a lobbyist with the Association of American Publishers.

The NIH is scheduled to submit a final version of the policy to Congress by 1 December, but Zerhouni says that the flood of comments makes it almost certain that this deadline will slip. ■





All together now: a proposed children's study will look at how biology and the environment interact.

## Report censures political screening of advisory boards

Geoff Brumfiel, Washington

Asking scientists questions about their political affiliations before allowing them to serve on government panels is "inappropriate", according to a report from the US National Academies.

The report, issued on 17 November, follows accusations by scientists and watchdog groups earlier this year that the Bush administration had politicized the process of appointing scientific advisory boards. "It is inappropriate to ask [scientists] to provide nonrelevant information, such as voting record, political-party affiliation, or position on particular policies," the report states.

"Clearly this report validates the concern expressed by senior scientists starting last winter about some of these issues," says Alden Meyer, director of strategy and policy at the Union of Concerned Scientists (UCS) in Washington DC. In February, the UCS accused the Bush administration of skewing scientific advice and screening committee members based on political beliefs (see *Nature* 427, 663; 2004).

Those allegations were viewed by Republicans as a political act, coming as they did in an election year. But Meyer says he believes the new study lends validity to the UCS claims. "Hopefully, having a committee of this sort come out after the election will carry some weight," he says.

Others are less sanguine about the academies' findings. "I think the report is a little naive," says Vernon Ehlers (Republican, Michigan), a nuclear physicist now serving in the House of Representatives. Ehlers says that although some panels address strictly scientific issues, many scientific advisory groups must wrestle with political issues. In those cases, Ehlers argues, "the president has a right to expect people to be in tune with his policies".

John Marburger, the president's science adviser, agrees that there is "some ambiguity" in the report over how to deal with advisory committees handling highly political topics. But he says the report gives a good overall view of the advisory committee process.

Marburger adds that the use of political questions to screen candidates has been "fairly rare". Still, he says, "if anyone feels they're being asked inappropriate questions in connection with membership on a science advisory committee, I sure would like to know about it".

## Huge study of children aims to get the dirt on development

Erika Check, Washington

The architects of a study that would follow 100,000 American children from conception to adulthood have unveiled their plans after four years of preparation.

The designers of the National Children's Study, which would cost \$2.7 billion to run, hope to collect a wealth of information. Sources will range from blood taken from mothers before pregnancy to samples of the dirt found in children's homes, schools and playgrounds. The study blueprint, released on 16 November, also calls for information to be collected on the children's genes, chemicals in their bodies and the structure of their families.

The data will shed light on the broad issue of how biology and the environment interact to cause diseases and developmental disorders. But the study team, which is based at the National Institute of Child Health and Human Development in Rockville, Maryland, is particularly interested in areas such as obesity, mental disorders and issues involving pregnancy, including birth weight and birth defects.

"There's a lot of hope pinned on this," says study adviser Nancy Green, medical director of the March of Dimes in White Plains, New York, which lobbies on health issues in babies. Green says the study could be as crucial as the famous Framingham Heart Study, which has followed residents of a town

in Massachusetts since 1948 and is credited with having revolutionized the treatment of cardiovascular disease.

The study could also get to the roots of rare disorders, such as types of childhood cancer, by sharing data with similar efforts in other countries. Investigators in Scandinavia, for example, this year began collecting biological samples and data on 200,000 Danish and Norwegian children, although they are not directly sampling the children's environment.

The child health institute, which was given \$50 million of government funding in 2000 to plan the study, now needs to persuade Congress to fund the project. It has the backing of a diverse group of 48 organizations, including paediatric health groups and the industry-funded American Chemistry Council, which announced its support in a letter to the institute on 12 November. "The cost of the study is dwarfed by the cost of treating the diseases and conditions it can be expected to address," the letter states.

The March of Dimes director of public policy and government affairs, Jo Merrill, hopes that such arguments will convince Congress to support the child study, even though spending on domestic programmes has been drastically curtailed because of the US war on terrorism. "We're cautiously optimistic," she says.

## Junior scientists are denied access to data, says survey

Rex Dalton, San Diego

A quarter of trainee scientists have had requests for data denied, according to a survey of more than 1,000 researchers at 50 of the largest research institutions in the United States.

The study, which is the first attempt to estimate the extent to which doctoral students and postdoctoral fellows suffer from data-withholding, found that 23% of respondents were denied results, materials or other information related to published research. Of the three disciplines surveyed, life sciences (27%) had greater difficulties than chemical engineering and computer science (20%).

Withheld data delayed respondents' research by about four months on average, says Eric Campbell, a sociologist at Harvard University who discussed the results on 12 November at the Office of Research Integrity's conference in San Diego. Campbell and his colleagues have charted data-access issues for several years (see *Nature* 404, 6; 2000), but this is the first study to focus on PhD students and postdocs.

Campbell did not discuss why the data were being withheld, but said that he had considered and rejected various explanations during the study. Restrictions due to medical privacy or industrial contracts, for instance, could not account for the pattern of data-withholding observed, Campbell says.

The survey, which was sponsored by the Office of Research Integrity, could help raise awareness of the data problem. When the study began early last year, Campbell says that one university department chairman told him: "Trainees don't know anything about data being withheld from them; why would you want to ask them?" ■



Raw deal: Eric Campbell says data are withheld from one in four young biologists.



Hard cell: tissue banks store vast quantities of material but organize it in different ways.

## Summit calls for clear view of deposits in all biobanks

Helen Pearson, New York

Action must be taken to coordinate human-tissue banks if these resources are to be exploited to the full, researchers said at a meeting last week.

In recent years there has been a huge increase in the number of biobanks — collections of tissue or DNA samples matched with health records. The samples should allow disease susceptibilities and drug responses to be linked to variations in genes or proteins. Yet little has been done to maximize the banks' potential by ensuring information can be shared between them, say researchers who met at the Biobank Summit in Tarrytown, New York, over 15–17 November.

"All the effort and immense public money will be wasted if we can't combine certain elements," says Bartha Knoppers, a bioethicist at the University of Montreal, Canada.

Problems with biobanks will arise, researchers say, because each one tends to have a different system for collecting, processing and storing tissue, logging medical histories and obtaining patient consent. They also use different computer systems and codes to store information.

These differences could hamper a scientist who wants to pool sufficient patients to study a rare cancer, for example. Tissue samples in two biobanks may be impossible to compare if they have been processed using different chemicals. And medical histories could be worthless if they use differing terminology or lack essential information about diet or lifestyle.

Some kind of international coordination is "extremely important if we are ever going

to understand health on a big scale", says Jan-Eric Litton, director of informatics at the Karolinska Institute Biobank in Stockholm, Sweden.

A few collaborations are emerging. Last year, Knoppers and others started the Public Population Project in Genomics, which aims to build a database of reference documents on best practices for collecting samples, medical histories and consent. The project received seed funding last week of Can\$1 million (US\$0.8 million) from Génome Québec and Genome Canada.

Another group is focusing on cancer biobanks around the world. It plans to meet in early 2005 to discuss how to ensure that information can be exchanged. "We have to act pretty soon on this," says Anna Barker of the National Cancer Institute in Bethesda, Maryland, who heads a programme to build a US cancer-patient biobank and is helping to organize the cancer group's meeting next year.

Such moves are urgent because of the plethora of biobanks springing up around the world. Britain, Estonia and Sweden are among the countries creating huge, government-sponsored biobanks. Countless more are being built up by academic groups, non-profit organizations and the biotechnology and pharmaceutical industries.

Yet most researchers have been too focused on establishing their own biobanks to think much about linking up with others, delegates said at the meeting. "We will be playing catch-up," says pathologist Carolyn Compton of McGill University in Quebec, Canada, "but it is doable." ■



# Chemistry claim provokes strong reaction

**Alison Abbott, Munich**

A bitter dispute between Nobel laureates that has festered behind the scenes for nearly 40 years is now getting a very public airing.

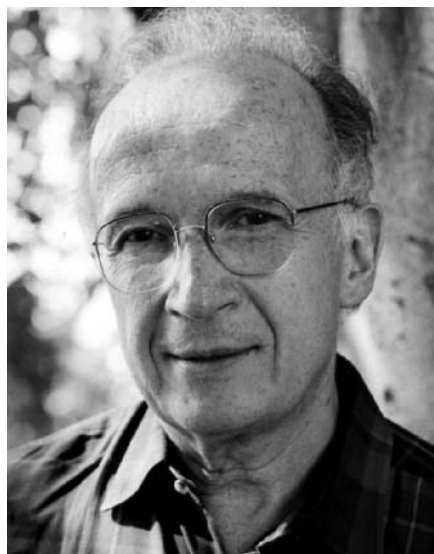
The argument centres on who should get the credit for a set of rules that help to predict the outcome of certain chemical reactions. One of the three chemists involved in the argument is taking the unusual step of publishing his version of events in a leading chemistry journal.

On 19 November, Roald Hoffmann of Cornell University in Ithaca, New York, winner of the 1981 Nobel Prize in Chemistry, put his side of the story online in *Angewandte Chemie* (R. Hoffmann *Angew. Chem. Int. Edn* **43**, doi:10.1002/anie.200461440; 2004).

The catalyst for this move was a claim made in public for the first time by E. J. Corey, a chemist at Harvard University, who won a Nobel in 1990. In an acceptance speech for the American Chemical Society's Priestley medal earlier this year, Corey said that he sowed the seed of the idea that led to Hoffmann's prize. The third chemist involved in the row, Robert Burns Woodward, received his Nobel prize in 1965 but died in 1979.

The story begins in the mid-1960s, when all three chemists were at Harvard. Woodward was working on the synthesis of vitamin B12, and was trying to find out why the molecular geometry of some reaction products was the opposite of that predicted by theory.

He believed that the answer might lie in quantum-mechanical effects related to the orbits of electrons in organic molecules, and he recruited Hoffmann — a junior fellow with a growing reputation as a talented theorist — to help him develop this idea. Together, they worked out the Woodward–Hoffmann



Nobel laureate Roald Hoffmann hopes to settle an argument over rules that bear his name.

rules, which predict the outcome of certain organic reactions based on the orbits of the electrons in the starting materials.

The pair established that the rules applied to a wide class of reactions in organic chemistry, and in doing this helped to revolutionize the field, says Henning Hopf, a synthetic chemist at the Technical University of Braunschweig in Germany.

But Corey, who knew both Woodward and Hoffmann, has now gone public with his claim that he first suggested the idea to Woodward. "On May 4, 1964, I suggested to my colleague R. B. Woodward a simple explanation ... that provided the basis for the further development of these ideas into what became known as the Woodward–Hoffmann rules,"

he said as part of his acceptance speech for the Priestley medal (Corey, E. J. *Chem. Eng. News* **82** (13), 42–44; 2004).

Corey had previously told several people, including Hoffmann, that he believed he had seeded the idea that led to the rules — and that he was offended that his input was not acknowledged. But he did not speak about this to Woodward, who received his Nobel for an unrelated idea.

In his article, Hoffmann argues that Corey's claim is "not right". He recalls Corey telling him of the 1964 conversation "some-time in the 1970s" and not taking it seriously. He adds that only days after his Nobel prize announcement in 1981, he received a letter from Corey repeating the claim. Corey asked Hoffmann to mention his contribution in his Nobel speech, which Hoffmann declined to do. "I found it unfair that Corey was asking me to make his claim for him," Hoffmann says. "Especially as Woodward had already died and could not comment."

Corey later told Hoffmann that he hadn't wanted to raise the subject himself because he didn't want to damage Harvard — and because he thought that Woodward would mellow with age and "grow more sensitive to his own conscience".

Scientists who were in the Corey and Woodward labs at the time say that they are embarrassed by the episode, and puzzled that Corey should reveal his claim after 40 years of silence. "E. J. probably did say something to Woodward about molecular orbital theory, but it was Woodward who picked up the baton and ran with it, with Hoffmann," says Ian Fleming, a chemist at the University of Cambridge, UK, who was a postdoc with Woodward in 1964. Corey declined to comment. ■

C. A. LOEW

## Asian nations build bridges to bolster science

**K. S. Jayaraman, New Delhi**

China and India have signalled a thaw in their previously frosty relations by agreeing to forge closer scientific ties, including plans to work on nuclear energy and space research.

The two countries, whose scientific productivity has mushroomed over the past decade, announced on 17 November that they will set up a Joint Steering Committee, co-chaired by their science ministers, to promote collaboration.

"This is the first time that cooperation in science and technology between the two nations has gone to ministerial level," says Valangiman Ramamurthi, India's science secretary. "Until now, collaboration has consisted of the exchange of scientists and holding workshops."



China is to work with India on nuclear projects, such as this reactor being built near Madras.

The committee, which will comprise top officials from government departments, will be decided in the next few weeks. A formal agreement will be signed when Indian science minister Kapil Sibal visits

Beijing early in 2005, Ramamurthi says.

The decision to create the committee capped a week-long tour of Indian scientific institutions by a 30-strong Chinese delegation led by state councillor Chen Zhili. The fact that Chen is the first Chinese minister to visit India is a measure of how seriously China views the collaboration, Indian officials say. The delegation was particularly keen to collaborate on biotechnology, space science, nuclear technology, oceanography and research in herbal medicines, says Sibal.

Ramamurthi adds that many things can be done jointly. Combining Chinese strength in electronic hardware with India's skills in software could, for example, make the pair a formidable force in the global information-technology market, he says. ■

M. LAKSHMAN/AP

## Wisconsin follows California in funding stem-cell research

**San Francisco** Wisconsin has decided to throw \$750 million into a plan to bolster stem-cell research. Observers say the move is likely to be the start of a string of state initiatives geared to preventing stem-cell researchers from fleeing to California, where a \$3-billion bond measure for embryonic-stem-cell research was approved on 2 November (see *Nature* **432**, 135; 2004). Both states are bucking federal policies that provide only limited funding for some stem-cell research.

"This is not a competition," governor Jim Doyle said in announcing the venture. "If anything, there will be synergy between our two states."

Researchers at the University of Wisconsin in Madison isolated the first human embryonic-stem-cell line in 1998, and the state has since become a leader in the field.

The governor's package, announced on 17 November, combines state and private funds to set up a new research institute that will subsume WiCell, the private foundation that currently distributes the university's stem-cell lines.



Burning issue: cars torched in pollution protest.

## Ecowarrior student faces five-year stretch for arson

**San Diego** A physics graduate student at the California Institute of Technology was convicted of arson on 19 November for torching sport utility vehicles during an Earth Liberation Front (ELF) crime spree last year. The ELF aims to damage those it deems to be exploiting the environment.

**William Cottrell, 24, who was studying string theory when arrested last March by federal agents, faces at least five years in prison after conviction in a US District Court in Los Angeles on seven counts of arson and one count of conspiracy to commit arson. The protests caused damage costing more than \$2 million. His sentencing is scheduled for 7 March 2005.**

Cottrell's lawyer says the student was

particularly vulnerable and prone to social manipulation because he has an autistic condition, Asperger's syndrome. He says he plans to appeal against the verdict.

The jury acquitted Cottrell of the more serious felony charge of using a destructive device during a violent crime, for which he could have earned 30 years in prison.

## Rejected physicists instigate anti-arXiv site

**Washington** Researchers who feel they have been unfairly excluded from the arXiv physics preprint server now have a new home on the Internet.

The 'archive freedom' site, developed by a handful of frustrated researchers, hosts the stories of physicists who, they claim, have been "blacklisted" by arXiv's operators at Cornell University in Ithaca, New York. The site includes information about Robert Gentry, a geophysicist formerly at Oak Ridge National Laboratory in Tennessee. Gentry, a Seventh Day Adventist and creationist, lost a legal action this March in which he had accused arXiv of religious discrimination in rejecting his papers on an alternative to the Big Bang theory (see *Nature* **428**, 458; 2004).

Paul Ginsparg, a physicist at Cornell who founded arXiv in 1991, defends the archive's policies and says the rules governing who



can and cannot publish are clearly stated on the site. The archive is not a fully open forum, he adds, and is designed for “communication among research professionals, not as a mechanism for outsiders to communicate to that community”.

## Soldier ‘unlawfully killed’ in 1950s nerve-gas tests

**London** A verdict has finally come in on the case of a British serviceman who died in 1953 after being deliberately exposed to the nerve gas sarin by army officials. On 15 November, a coroner conducting a second inquiry into the case announced that Ronald Maddison had been unlawfully killed.

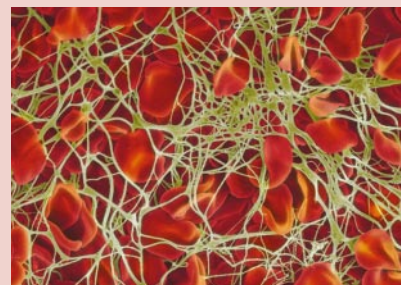
The decision goes against the original 1953 verdict of death by misadventure, and is a boost to veterans who have long campaigned for an investigation of experiments at the Ministry of Defence’s Porton Down facility in Wiltshire between 1939 and 1989. Maddison died aged 20 after sarin was dripped on to a patch of uniform taped to his arm. About 500 servicemen say that they were involved in related tests, and many say they suffer health problems as a result. A decision was made last year not to pursue criminal prosecution because of a lack of evidence.

## Labs align as standard is set for genetic test

**London** In an effort to bring results in genetic labs around the globe into line, the World Health Organization has approved the first international standard for a genetic test.

“It’s genetic testing coming of age,” says Stephen Inglis, director of the UK National Institute for Biological Standards and Control. The Hertfordshire-based institute developed the standard, and is now working on more.

The standard consists of DNA samples with and without a genetic mutation called factor V Leiden, which is linked to the formation of potentially fatal blood clots in the veins (pictured). Labs can now use the reference DNA to see exactly what tests in their own lab should



look like for both positive and negative results.

Experts say that the development of such reference standards will become ever more important as the number of genetic tests grows.

**With the Maddison decision in place, the Porton Down Veterans Support Group is now calling for an independent inquiry into the rest of the facility’s activities.**

## UK urged to admit Gulf War syndrome exists

**London** An independent inquiry has called for the UK government to acknowledge the existence of Gulf War syndrome.

The report concludes that about 6,000 British veterans of the 1991 Gulf War have

suffered ill health as a direct result of their service, and that veterans of this conflict are twice as likely to suffer from poor health as those who served elsewhere.

The inquiry, headed by Lord Lloyd of Berwick, was funded by an anonymous source.

Britain’s Ministry of Defence has always denied that Gulf War syndrome exists, although there is a special pension for veterans of this conflict. It is expected to be weeks before the government officially comments on the report.

# We have the technology

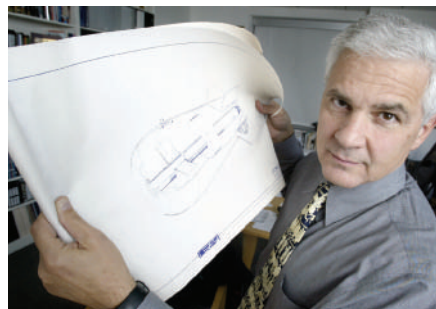
The global spread of nuclear weapons is once again a major headache for world leaders. Geoff Brumfiel reports on efforts to put the genie back in the bottle.

**“H**ere’s my bomb design,” quips Joe Cirincione, unrolling a blueprint on his office table. The paper shows a cutaway view of a teardrop-shaped device with a flattened cylinder of uranium-235 at one end, and a small plug of the same highly enriched metal at the other. When the two pieces are brought together at the right speed, they begin a chain reaction with the explosive power of several thousand tonnes of TNT.

Cirincione, who directs the Carnegie Endowment for International Peace’s non-proliferation programme in Washington, bought the simplified design a few years ago from the gift shop at Los Alamos National Laboratory in New Mexico, where the world’s first nuclear-bomb programme began 60 years ago.

Picking up a real, detailed bomb blueprint is not yet quite as simple as visiting a shop — but it seems to be much easier than experts thought just a few months ago. In January, Abdul Qadeer Khan, the father of Pakistan’s nuclear bomb, confessed to heading an extensive network of scientists, engineers and businessmen who were selling nuclear secrets on the black market. The group’s bill-of-fare included complete design data for at least one tested nuclear warhead.

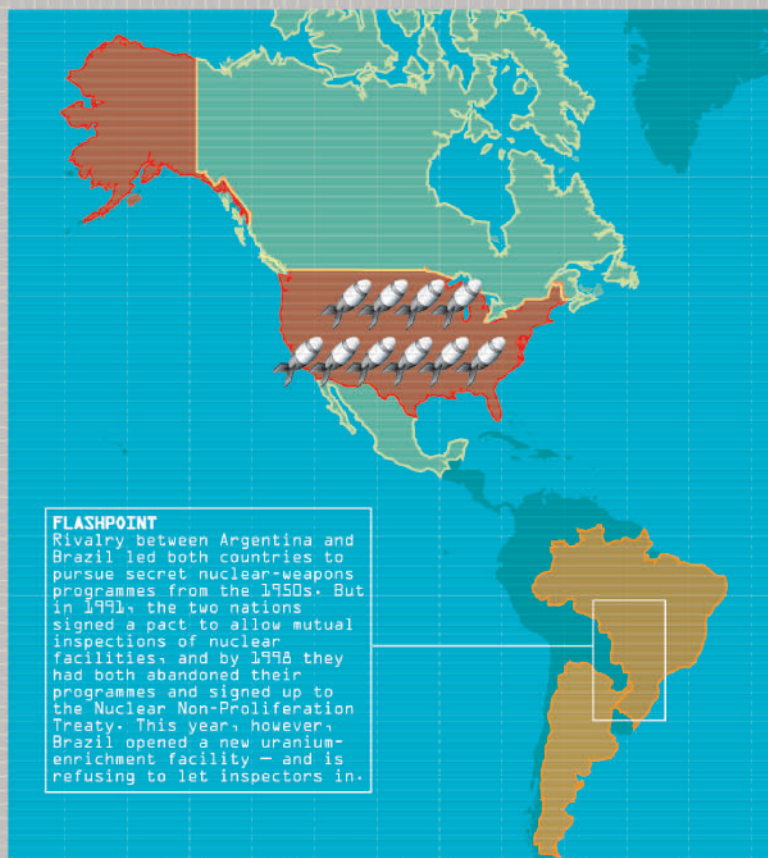
The revelation that a private network was selling such a blueprint was a wake-up call for many politicians and arms-control experts, says Corey Gay Hinderstein, deputy



Joe Cirincione shows off his souvenir bomb plan.

## NUCLEAR WEAPONS UPDATE

Tonnes of plutonium  
Tonnes of highly enriched uranium  
Number of nuclear weapons



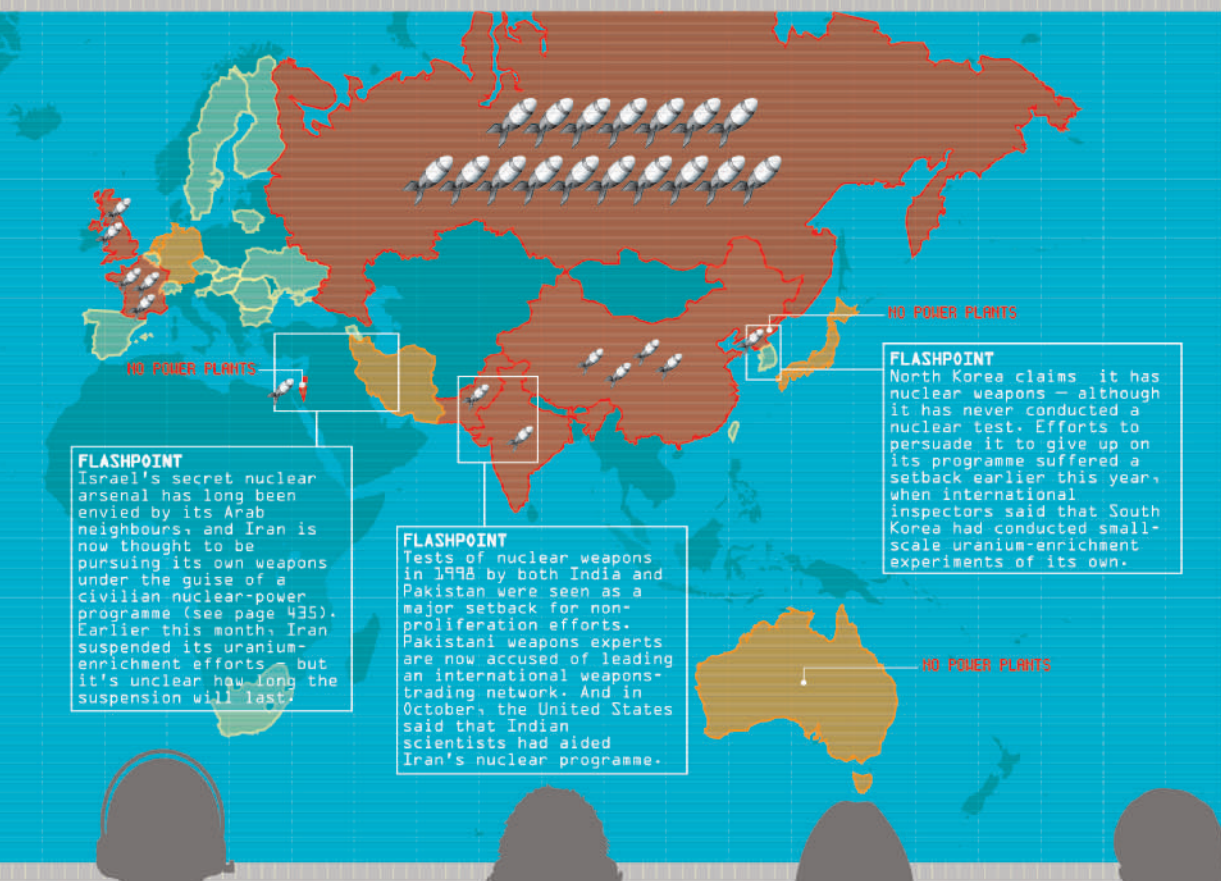
### Countries with:

- Nuclear power plants
- Nuclear power plants + Uranium-enrichment or plutonium-reprocessing plants
- Nuclear power plants + Uranium-enrichment or plutonium-reprocessing plants + Nuclear weapons

\*\*\*\*\* ISRAEL NUCLEAR ON

| US     | Russia    | China | France | UK  | Israel | India | Pakistan | N. Korea |
|--------|-----------|-------|--------|-----|--------|-------|----------|----------|
| 100    | 150       | 5     | 5      | 8   | 0.5    | 0.5   | 0.04     | 0.02     |
| 700    | 800-1,400 | 20    | 30     | 20  | ?      | ?     | 1        | ?        |
| 10,000 | 17,000    | 400   | 350    | 200 | 100    | 70    | 45       | 0-5      |

All figures are approximate and rounded



SOURCE: INSTITUTE FOR SCIENCE AND INTERNATIONAL SECURITY/CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE/INTERNATIONAL ATOMIC ENERGY AGENCY





**Time bomb:** the United States' nuclear arsenal is envied by a clutch of aspirant nuclear powers.

director at the Institute for Science and International Security, a Washington-based think-tank. "We're used to states putting together these efforts," she says. "But this was pretty amazing."

The Khan network is just part of a rapidly changing nuclear landscape. The end of the cold war has left a glut of fissile materials in Russia that could potentially be fashioned into a crude bomb by terrorists or sold on the black market to states with nuclear ambitions (see 'Russia: under lock and key', page 436). Countries such as South Korea and Brazil, which once might have struggled to develop nuclear weapons on their own, now have the technical ability to do so if they wish. And Iran and North Korea, two nations that signed an international treaty meant to stop the spread of nuclear weapons, either have bombs of their own or are very close to having them (see 'Nuclear weapons update', previous page).

### On the brink

Whether these developments will unravel 60 years of effort to control the nuclear-arms race is open to question. But one thing seems clear — old assumptions about who can get a nuclear weapon no longer hold. "We are at a nuclear tipping point," Cirincione says. "The decisions that we make over the next couple of years will decide whether progress continues or whether we go off on the second great proliferation wave since the Second World War."

Proliferation was a deep concern after 1945. With tensions between the Soviet

Union and the United States rapidly on the rise, many nations began to develop nuclear programmes as a safeguard against an uncertain future. The threat was all too clear to President John F. Kennedy, who narrowly avoided nuclear war when the Soviet Union tried to place nuclear missiles on Cuba in October 1962. "I ask you to stop and think for a moment what it would mean to have nuclear weapons in so many hands. ... There would be no rest for anyone then, no stability, no real security, and no chance of effective disarmament," he said in July 1963 in a televised speech that announced the first major US–Soviet treaty, which banned atmospheric testing of nuclear weapons.

By the end of the 1960s, the United States and the Soviet Union had signed the Nuclear Non-Proliferation Treaty (NPT), a landmark agreement that defined the next 30 years of nuclear-arms control. The treaty pledged that states with nuclear weapons would prevent the spread of weapons technology to non-nuclear nations — and that they themselves would work towards complete nuclear disarmament. Non-nuclear nations agreed not to develop weapons as long as all countries would have access to peaceful nuclear technology, including nuclear power. The treaty also specified that states without

Union and the United States rapidly on the rise, many nations began to develop nuclear programmes as a safeguard against an uncertain future. The threat was all too clear to President John F. Kennedy, who narrowly avoided nuclear war when the Soviet Union tried to place nuclear missiles on Cuba in October 1962. "I ask you to stop and think for a moment what it would mean to have nuclear weapons in so many hands. ... There would be no rest for anyone then, no stability, no real security, and no chance of effective disarmament," he said in July 1963 in a televised speech that announced the first major US–Soviet treaty, which banned atmospheric testing of nuclear weapons.

By the end of the 1960s, the United States and the Soviet Union had signed the Nuclear Non-Proliferation Treaty (NPT), a landmark agreement that defined the next 30 years of nuclear-arms control. The treaty pledged that states with nuclear weapons would prevent the spread of weapons technology to non-nuclear nations — and that they themselves would work towards complete nuclear disarmament. Non-nuclear nations agreed not to develop weapons as long as all countries would have access to peaceful nuclear technology, including nuclear power. The treaty also specified that states without

nuclear weapons must comply with rules set up by the International Atomic Energy Agency (IAEA).

In addition to the diplomatic barrier formed by the NPT, would-be nuclear-weapons states faced formidable technical hurdles. Like all forms of nuclear energy, bombs get their power from the massive amounts of energy released when trillions of atomic nuclei are split apart. The bombs that are easiest to build — such as the design in Cirincione's office — rely on uranium-235, a relatively rare isotope of the metal found in small quantities in ore deposits.

### Enriched pickings

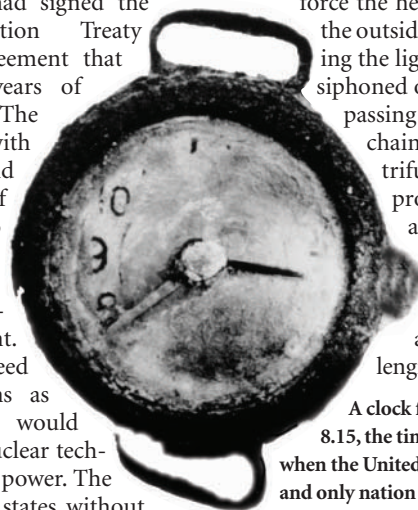
The bulk of natural uranium is the heavier uranium-238 isotope, so to produce fuel for a nuclear reactor or a bomb the 235 isotope must be separated from the 238. When the proportion of uranium-235 in the metal reaches about 5%, it can sustain a controlled fission reaction in a nuclear-power plant. A much higher level of uranium enrichment — 90% or more uranium-235 — will produce a nuclear explosion when the metal is brought together in a 'critical mass' of just a few kilograms.

The two uranium isotopes are chemically identical, and can only be separated using techniques that exploit the difference in their atomic masses. Until the 1960s, this meant using gas-diffusion facilities, which separate the atoms by taking advantage of the slightly different rate at which they pass through a maze of chambers and membranes. These immense facilities were difficult for aspiring weapons states to build and easy for inspectors to spot, says Ernest Moniz, a theoretical physicist at the Massachusetts Institute of Technology and former under-secretary of the US Department of Energy, which runs the US nuclear-weapons programme. But in the 1970s Urenco, a European consortium for making nuclear fuel, developed another concept that would prove far easier to hide.

Urenco used centrifuges, which spin gaseous uranium fluoride at high speeds to force the heavier uranium-238 to the outside of the chamber, leaving the lighter 235 isotope to be siphoned off from the inside. By passing the gas through a chain of thousands of centrifuges, it is possible to produce nuclear fuel and, with a little more effort, bomb-grade material.

The centrifuges are technically challenging to build. They need

A clock from Hiroshima still says 8.15, the time on 6 August 1945 when the United States became the first and only nation to use nuclear weapons.



## Middle East: politics and power plays

Simple phrases can have far-reaching power. When the US foreign-policy establishment coined the term 'rogue state' in the mid-1990s, its members were well aware of this. The phrase, which is often used as an epithet for nations that seek nuclear weapons, triggers images of maverick, unstable governments whose leaders want the bomb at any cost, and don't care about the consequences.

Yet when nations choose to pursue nuclear-weapons programmes, the decision usually reflects the cold demands of local politics. This is the case even in regions where the spread of nuclear weapons is sometimes characterized as scary and unpredictable, such as the Middle East, where Israel and Iran have each fostered undeclared nuclear-weapons research programmes.

The history of Israel's clandestine programme, which began shortly after the establishment of the Jewish state in 1948, reads like a Tom Clancy novel. Parts of the country's fledgling nuclear facilities at Dimona, in the Negev Desert of southern Israel, were probably hidden from inspectors by bricking up elevators and installing false control panels. Weapons may have been secretly tested underwater in the Indian Ocean, when cloud cover was shielding the bright explosions from watching satellites.

### Calculated risk

The country is now believed to have about 100 warheads. Details are impossible to confirm as the programme does not officially exist, but Avner Cohen, a nuclear-proliferation expert at the University of Maryland in College Park, says that Israel could probably launch warheads from submarines, planes and by using missiles.

Initially, the programme risked angering Israel's closest ally, the United States, especially because it was US inspectors who were deceived when visiting Dimona. But Israel thinks the gamble was worthwhile. Nuclear weapons may have played a role in dissuading Israel's neighbours, in particular Egypt, from pursuing the path of

war, says Cohen. In the case of the 1973 Yom Kippur war with Syria and Egypt, he says, Israel may even have secured US military aid more rapidly because it threatened to use its nuclear capability.

Analogous calculations may now be going on in the minds of Iran's rulers. Officially, the country's nuclear programme is a civilian one. But since the 2002 announcement that Iran would build six nuclear power stations, the International Atomic Energy Agency (IAEA) has accused the nation of misreporting the amount of uranium it has imported, and what it is doing with it. Iran is entitled to own the uranium as part of its power plans, but most Western analysts think that it provided misinformation in a bid to hide a weapons programme.

The programme, which is dispersed around at least five locations, is thought to include efforts to produce both plutonium and highly enriched uranium.

According to the Carnegie Endowment for International Peace, a Washington-based think-tank, one of the most advanced facilities is at Natanz, some 300 kilometres south of Tehran. Iran says that the site, which has been visited by the IAEA, will produce reactor fuel. But when running at full capacity, it could make 400–500 kilograms of weapons-grade uranium, enough for 15–20 weapons, per year.

### In the dock

The United States, which already imposes sanctions against Iran for its alleged links to terrorism, wants the country referred to the United Nations Security Council over the IAEA's assessment of its uranium imports. Any weapons programme would breach the Nuclear Non-Proliferation Treaty, which Iran signed in 1968. By violating the treaty, Iran could

alienate nations with which it has less fraught relations, such as France and Germany.

Yet from a national-security perspective, Iran might still decide that a nuclear-weapons programme is worthwhile. Two of Iran's immediate neighbours, Iraq and Afghanistan, are now occupied by the United States.

Israel's nuclear weapons can probably reach Iranian territory. And Iran must also consider whether a revitalized, independent Iraq — should one emerge — would resume old hostilities.

"Over time, Iran will create the impression that it is close to developing the bomb," predicts Cohen. "That could have a deterrent effect."

Internal Iranian politics certainly point in this direction. The country has looked set for change since reformists won elections in 1997, but they have lost some popular support after struggling to overcome resistance from conservative religious groups. The idea that Iran has the right to nuclear energy and nuclear weapons unites supporters of both factions. "The public-pride issue is huge," says Michael Levi, an arms-control expert at the Brookings Institution in Washington.

If Iran continues to defy the West and develops nuclear weapons, drastic action seems unlikely, at least in the short term. Israel ended Iraq's nuclear programme in 1981 by bombing the country's reactors, but Iran's facilities are farther away and more dispersed. And even hawks in the United States are reluctant to confront Iran — which is nearly three times more populous than Iraq — militarily.

European nations have had some success in holding back Iran's nuclear ambitions by offering trade and aid packages. But Iran has so far said only that it will suspend, not halt, its programme. And if it continues to accrue nuclear technology, experts fear yet more proliferation in the region, with Saudi Arabia, Syria or Egypt next in line.

**"Over time, Iran will create the impression that it is close to developing the bomb. That could have a deterrent effect."**

— Avner Cohen



Clandestine operations: civilian nuclear sites in both Iran (top) and Israel are believed to harbour production of weapons-grade material.

EPA/PA PHOTOS

GETTY IMAGES

Jim Giles



## Russia: under lock and key

Top Russian officials no longer pull their punches when describing the prospect of nuclear proliferation on their doorstep.

"We cannot fully rule out the probability of Russian fissile materials, as well as technologies lending themselves to the creation of nuclear weapons, falling into the hands of terrorists," Alexander Rumyantsev, Russia's minister of atomic energy, told a September meeting of the Global Threat Reduction Initiative in Vienna.

On the ground in Moscow, researchers at the birthplace of the Russian atomic bomb take a more sanguine view of security at their own facility.

"Efforts to improve the physical protection of weapons material originated here ten years ago," says Nikolai Ponomarev-Stepnoi, vice-president of the Kurchatov Institute, Russia's premier institute for basic research into weapons physics and fission technology. "Our own activities have become the impetus for security upgrades at many locations throughout Russia."

The ageing, but well-maintained 100-hectare complex in Moscow houses the world's oldest nuclear reactor outside the United States. The reactor first attained fission in 1946 and still serves as a neutron source. To this day, all kinds of fissile material, including substantial amounts of weapons-grade enriched uranium and small amounts of plutonium, are stored here behind brick and corrugated iron.

### Under control

Some 5,000 staff still work at the institute, located only twelve kilometres from the Kremlin, and now surrounded by restless Moscow traffic and overlooked by dreary concrete apartment blocks.

Ponomarev-Stepnoi greets visitors in a chilly, low-ceilinged conference room on the edge of the site, and reassures them that here, at least, fissile materials are under careful control.

"We understand very well that we must continuously work on risk reduction," he says. "We also know that we can't solve the problem



The remnants of Russia's nuclear programme (top) are kept behind bars at places such as the Kurchatov Institute.

alone. But thanks to US Department of Energy support, the Kurchatov Institute has become one of the best-protected nuclear sites in Russia."

From 1949, when it carried out its first test, until the mid-1980s, the Soviet Union built about 45,000 nuclear weapons. This cold war arsenal has been heavily reduced since. In the past 15 years, Russia has dismantled thousands of tactical weapons, and in 2002 it agreed to cut its number of strategic warheads to between 1,700 and 2,200 by the end of 2012.

But despite substantial disarmament, the nuclear legacy of the Soviet Union remains the most severe headache for non-proliferation experts worldwide. Russia still has a massive nuclear-weapons production complex, including some 3,500 weapons scientists, roughly 100 of whom are believed to have comprehensive

knowledge of bomb design. Weapons-grade material is stored at about 250 sites of variable vulnerability, stretching from Murmansk to Vladivostok.

### Shut in

Perhaps the greatest risk is that insiders might try to sell valuable material to criminal groups. But nuclear-watchdog groups think that genuine progress has been made in improving the protection of the former Soviet nuclear facilities.

Since 1994, roughly half of the 250 sites have received some kind of security upgrade under the US-funded Material Protection, Control, and Accounting programme. These range in scope from rapid measures, such as bricking up windows, to high-tech security systems such as those installed at the Kurchatov Institute.

But *Securing the Bomb*, a recent analysis commissioned by the Nuclear Threat Initiative (NTI), a US watchdog group, concluded that truly 'comprehensive' upgrades have been completed at only one-fifth of all Russian sites storing weapons material, and at merely 5% of about 150 military sites believed to house actual warheads.

These upgrades have slowed lately, as the administration of President Vladimir Putin has become more reluctant to allow

foreigners into sensitive Russian facilities.

At the ten-year anniversary meeting last month of the International Science and Technology Center (ISTC) — a multilateral agency that redirects activities formerly related to weapons research into civilian projects — experts stressed the urgency of reinforcing joint non-proliferation efforts.

"At the current rate, it will take at least another decade to sufficiently secure all nuclear material in Russia," says Charles Curtis, president of the NTI. "We must do a lot more, a lot faster."

Since 1992, ISTC grants have helped thousands of Soviet weapons scientists survive in the market economy. Although not all Russian officials were enthusiastic about the programme in the first place — there were widespread concerns about national knowledge being sold off on the cheap — the success of ISTC-funded projects, from environmental research to software development, has convinced many sceptics.

Ex-Soviet weapons scientists are not now thought to be particularly susceptible to illicit offers to buy their knowledge.

Extra grants, such as those from the ISTC, allow most of them to maintain a decent standard of living. But badly paid guards, technicians and machinists working at Russian nuclear sites don't get similar aid, and fears remain that

some of them could be tempted by financial offers.

In the age of the terrorist warfare, the Soviet Union's nuclear arsenal has become a national security threat to Russia itself, and non-proliferation efforts have become a military necessity. "Our task is to always be one step ahead of terrorists," says Ponomarev-Stepnoi. "I don't believe that they will ever be able to get hold of a Russian weapon. But if somebody says there is no risk, everyone will go to sleep."

Quirin Schiermeier



powerful motors, strong and lightweight materials to withstand the stresses of the spinning, and special bearings to house the chambers. But they require significantly less space and power than a gas-diffusion plant, making them more efficient and easier to conceal. These traits make centrifuges “the proliferation technology *du jour*”, Moniz says.

Since its debut, centrifuge technology has gradually spread across the globe. Some non-weapons states, including Japan and Germany, use it for their nuclear-power industries, as the NPT permits. In other cases, centrifuge technology has percolated through illicit channels. From 1972 to 1975, Khan, for example, worked for a firm collaborating with Urenco, where, according to arms-control experts, he stole centrifuge designs that provided the basis for the nuclear weapon that Pakistan successfully tested in 1998.

### Secrets for sale

Later, Khan offered the centrifuge designs to Iran, Iraq, Libya and North Korea, among others. He turned the international network he had built for Pakistan's bomb programme into a one-stop-shop for nuclear technology, according to Hinderstein. It manufactured components that weren't readily available on the open market and even offered whole centrifuges for sale. Among the items the network sold were plans for an outdated but functional Chinese warhead that could deliver tens of kilotonnes of explosive energy and was small enough to fit on a missile.

The assistance Khan offered, together with the availability of modern computers, precision machining equipment and advanced alloys, brought countries that a generation ago would have been considered too primitive to build their own bombs to the brink of nuclear statehood.

For US analysts, the most troubling of Khan's clients is Iran. Iran has ample reason to develop a nuclear bomb (see ‘Middle East: politics and power plays’, page 435), but the government says that it obtained the centrifuge technology to further its domestic nuclear-power programme — something that is perfectly legal under the NPT.

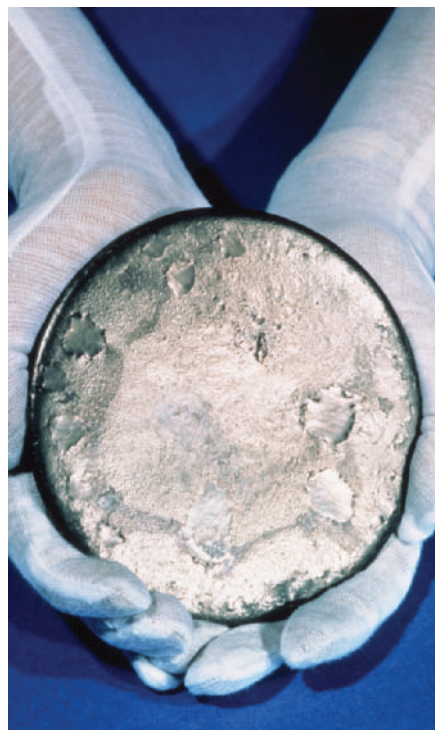
Paul Leventhal, head of the Nuclear Control Institute, a non-profit watchdog in Washington, says that Iran is taking the path followed by one of the United States' most steadfast allies: Japan. Since its first nuclear-power plant became operational in 1966, Japan has developed a large civilian nuclear programme that has produced several tonnes of plutonium-239, the other metal commonly used for nuclear bombs. Leventhal says that many consider Japan to be little more than “a screwdriver away” from a nuclear weapon. “Most think it could get a bomb in a matter of weeks to months, if not days,” Leventhal says.

It's a strategy that Hinderstein describes as “virtual proliferation”, and it seems to be catching on. Earlier this year, Brazil began operating a centrifuge-based uranium-enrichment plant that it says will provide fuel for civilian and naval reactors, but which could just as easily lay the foundation for a bomb programme. The nation has also limited

IAEA inspectors' visits to the facility. And in South Korea, a country long alarmed by the nuclear ambitions of its northern neighbour, a group of scientists recently admitted to

enriching a small amount of uranium using lasers — a technique that they had been developing for commercial purposes such as medicine and industrial testing. Laser separation of the different kinds of uranium would never be profitable for producing fuel, says Moniz — but it could be scaled up to make enough material for a handful of bombs.

Arms-control specialists say that virtual proliferation isn't really constrained by the NPT. The treaty was written at a time when nuclear power seemed to be the solution to the world's energy problems, and it was considered important that all nations had access to it. The main problem with treaty, according to John Wolf, president of the Eisenhower Fellowships programme in Philadelphia and former assistant secretary for non-proliferation at the US Department of State, is that it allows all countries access to all nuclear technology — including fuel-cycle technology that could be used for making bombs. If these technologies continue to spread,



Raw material: this lump of highly enriched uranium could be used to make a bomb.

Wolf predicts, “the treaty will break down”.

An international conference to review the NPT is set for next spring, but there is no consensus on what it ought to achieve. Mohamed ElBaradei, director-general of the IAEA, has called for an international treaty that would place fresh restrictions on the export of sensitive nuclear technologies. “We must universalize the export control system, remove these loopholes, and enact binding, treaty-based controls — while preserving the rights of all states to peaceful nuclear technology,” he wrote in *The New York Times* earlier this year.

### Going it alone

But the US administration is at loggerheads with ElBaradei and seems to set little stock by reinvigorating the NPT. Treaties are not the most attractive strategy at the moment, says Linton Brooks, head of the National Nuclear Security Administration, the arm of the US Department of Energy responsible for the nation's nuclear stockpile and many of its non-proliferation programmes. Instead, Brooks says that the United States advocates unilateral action, such as a moratorium on the sale of uranium-enrichment technologies abroad. The United States has also launched a Proliferation Security Initiative, under which it works with its allies to block trafficking of sensitive technologies.

A lively debate is unfolding in the broader arms-control community on how to proceed. Ted Carpenter, vice-president for defence and foreign-policy studies at the libertarian Cato Institute in Washington, argues that allowing nuclear weapons to spread will ultimately lead to stability.

Others seek to strengthen the NPT through new efforts to combat trafficking. At its annual meeting in June, the Carnegie Endowment, for example, unveiled a plan to reinforce the principles of the NPT by securing all nuclear materials, stopping illegal transfers, and committing nuclear states to regional conflict resolution (see *Nature* 430, 6; 2004).

Paul Robinson, director of Sandia National Laboratories in Albuquerque, New Mexico, thinks that a network of regional security alliances similar to NATO could watch over the world's nuclear weapons (see page 441).

Whatever the solution, Cirincione remains resolutely upbeat about the future. Twenty-five years ago, the world sat on the perpetual brink of nuclear annihilation at the hands of two superpowers, he says. Now it need only worry about regional nuclear war and the destruction of individual cities at the hands of nuclear terrorists. That's cause for some optimism, he contends: “The good news is that we're down to a few hard cases. The bad news is that they're really hard.” ■

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

# Replacement therapy, not recreational tonic

Testosterone, widely used as a lifestyle drug, is a medicine and should be kept as such.

Sir—We read with great interest the News Feature “A dangerous elixir?” (*Nature* **431**, 500–501; 2004) reporting the zeal with which testosterone is being requested by men who apparently view the hormone as a ‘natural’ alternative to the drug Viagra.

Like Viagra, testosterone replacement therapy — with emphasis on the word ‘replacement’ — is a solution to a very real clinical problem. There is a distinction between patients whose testosterone has declined to abnormally low levels, and individuals seeking testosterone as a pick-me-up. (As clinical and basic researchers interested in the effects of testosterone on the cardiovascular system, we find that many clinicians and scientists also seem unable to make this distinction.) Perhaps it is not surprising that testosterone is turning into a lifestyle drug.

During the Depression, the United States was rife with travelling ‘doctors’ peddling

male tonics and remedies — most harmless, but some as drastic as grafting goat testicles onto humans. Unfortunately, a certain amount of quackery is now perceived to be associated with testosterone therapy, and the increased prescription of the hormone in recent years does little to dispel this notion.

Nonetheless, as your Feature highlights, hypogonadism is a very real clinical condition. It is characterized by abnormally low serum levels of testosterone, in conjunction with symptoms such as mood disturbance, depression, sexual dysfunction, decreased muscle mass and reduced bone-mineral density.

Heightened awareness of hypogonadism, together with the increasing incidence of associated conditions such as obesity, may in part explain the rise in testosterone usage. But the inappropriate prescribing of the hormone is likely to be a contributory factor.

We agree with the experts quoted in

your Feature that there is a need for large-scale, long-term clinical trials to address several issues associated with testosterone use. For example, testosterone therapy may be beneficial against osteoporosis, heart disease and Alzheimer’s disease, but the dangers remain obscure.

Until we know more, both prescribing clinicians and the male population need to be aware that testosterone, like any other medication, should only be administered to patients for whom such therapy is clinically indicated.

**Richard D. Jones\*, T. Hugh Jones\*,**

**Kevin S. Channer†**

\*Hormone & Vascular Biology Group, Academic Unit of Endocrinology, Division of Genomic Medicine, The University of Sheffield Medical School, Beech Hill Road, Sheffield S10 2RX, UK

†Faculty of Health & Well-being, Sheffield Hallam University, Collegiate Crescent, Sheffield S10 2BD, UK

## DDT still has a role in the fight against malaria

Sir—Your News story about the Roll Back Malaria campaign (“Struggling to make an impact” *Nature* **430**, 935; 2004) quotes me as claiming that pressure from government and other donors made spraying difficult to push through politically. I am also quoted as saying: “We have had very, very strong lobbying over DDT. We have had to give up.” The quotations give the impression that the World Health Organization (WHO) has given up on DDT under the pressure of lobbying. I believe this is misleading.

When interviewed, I explained that we sometimes had to give up trying to convince a specific donor to financially support indoor spraying with DDT, if they flatly refused because of its perceived toxicity and ecological hazard. This has occasionally occurred in countries where the government wished to use DDT, and there was evidence that it was the best option for malaria-vector control.

However, in general terms, the WHO has never given up in its efforts to ensure access to DDT where it is needed. At meetings of the intergovernmental negotiation committee on the Stockholm Convention — which seeks to control the spread of persistent organic pollutants — the WHO has successfully defended the right of countries to use DDT for disease-vector control, if no suitable alternative can be found. The WHO also supports worldwide efforts to develop alternative products and

phase in alternative control strategies ([www.pops.int/documents/meetings/inc4/en/inf9/inf9en.pdf](http://www.pops.int/documents/meetings/inc4/en/inf9/inf9en.pdf)).

The Stockholm Convention came into force in May this year. Its exemption allowing restricted and controlled use of DDT according to WHO guidelines is a good example of appropriate international regulation on a difficult dilemma. It is not a compromise but a solution, which ensures that disease-control programmes maintain access to a useful product, while fully respecting the need to prevent environmental damage from persistent organic pollutants, such as DDT.

**Allan Schapira**

Strategy and Policy Team, Roll Back Malaria Department, World Health Organization, Avenue Appia 20, 1211 Geneva 27, Switzerland

## Presidential candidates failed peer-review test

Sir—Peer review has had a bad press recently, but your piece by George W. Bush and John Kerry (*Nature* **431**, 240–243; 2004) shows how valuable it can be.

The article is conspicuously full of unsupported (and unsupportable) claims and deliberately ambiguous statements of the sort that never usually appear in *Nature*, because no reviewer would allow them. Maybe all statements by politicians should be refereed by scientists.

Of course, this idea also exposes the weakness of peer review. The article would

have to be refereed by Bush and Kerry’s peers, that is to say disingenuous politicians. Such referees would no doubt insist on removing any shred of clarity and simple truthfulness as contrary to accepted practice.

The principle of peer review is an excellent one, but the practice is only as good as the reviewers.

**Robert Insall**

School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

## Benveniste’s reputation was not written in water

Sir—The headline on your News story about the late Jacques Benveniste — “‘Memory of water’ biologist dies after heart surgery” (*Nature* **431**, 729; 2004) — reminds readers of the most controversial aspects of his career. His work cannot be fully appreciated by exploring only the second half of his scientific life. We should also remember his active participation in the discovery of platelet-activating factor, a potent pro-inflammatory lipid-derived mediator (J. Benveniste, P. M. Henson and C. G. Cochrane *J. Exp. Med.* **136**, 1356–1377, 1972; and J. Benveniste *Nature* **249**, 581–582, 1974).

Without omitting the more questionable part of his career, he deserves wider attention for this early work. His memory lives on.

**Bernard Rothhut**

CNRS UMR 6198, UFR Sciences, Université de Reims Champagne-Ardenne, 51687 Reims cedex 2, France



# Revisiting the Baruch Plan

Developing a realistic strategy to control the proliferation of nuclear arms.

**C. Paul Robinson**

Will we ever be able to banish the nightmare of nuclear war? With an increasing number of nations threatening to 'go nuclear', it is easy to conclude that the world is destined to live with the threat forever. Yet more than 50 years ago, the US statesman Bernard Baruch presented the world with a sober, well-thought-out solution. It may be wise to reconsider his ideas in the context of today's world of nuclear proliferation.

It is possible to be optimistic even when we consider the difficulty that has surrounded our attempts to curb the proliferation of nuclear weapons. In retrospect, it is remarkable that the nuclear devastation of Nagasaki and Hiroshima in 1945 — and the subsequent development of far more powerful weapons — did not generate a lasting international conviction that such weapons should never be used again. Sadly, the victors in that war did not immediately focus on either 'learning to live with the bomb' or preventing its further use. After the uneasy alliance of the former Soviet Union with the West collapsed and Moscow successfully tested its own bomb, there was a scramble to seek and maintain the nuclear high ground.

## An idealistic start

It was not so at the very beginning. In the Acheson–Lilienthal report of late 1945, Robert Oppenheimer and others suggested turning over nuclear-weapon work to an international agency. As a result, US representative Bernard Baruch went before the first session of the Atomic Energy Commission on 14 June 1946 and proposed a radical plan that would give the United Nations oversight of atomic weapons and nuclear power.

The Baruch Plan marked the culmination of several international 'summit' meetings that were held immediately after the end of the Second World War, each seeking to prevent warfare and the further use of nuclear weapons. Such ideas generated a fleeting euphoria, originating from the hope that this war had been so terrible that the international community would be prepared to go to extreme lengths



Bernard Baruch (right) warned that fear alone would not prevent the rise of nuclear weapons.

to prevent another. This vision is not dead, but slumbering fitfully after the trauma of the cold war.

In the Baruch Plan, the United States was prepared to hand over its atomic monopoly to a new UN Atomic Development Authority. The ultimate goal was the international control of nuclear research and the elimination of nuclear weapons.

**"I strongly believe that the advantage of nuclear deterrence reached a peak with the end of the cold war more than a decade ago, and has been fading ever since."**

"Terror is not enough to inhibit the use of the atomic bomb," Baruch said. "The terror created by weapons has never stopped man from employing them. For each new weapon a defence has been produced, in time."

The Baruch Plan proposals might have gone farther had it not been for the fact that the former Soviet Union was developing its own nuclear weapon capability. For this reason, the plan was not accepted, and eventually, a UN

organization with considerably less power, the International Atomic Energy Agency (IAEA), was vested with limited powers of supervision over nuclear activities.

The former Soviet Union tested its first nuclear weapon in 1949, followed in short order by Britain, France and China. Once the nuclear genie was out of the bottle, the idea began to emerge that — in a world of perpetual vulnerability to nuclear attack — the best hope was to find ways that would restrain any nation state from contemplating deliberate and major war. From this was born the concept of nuclear deterrence. The idea emerged slowly, as a 'derived truth' rather than from any thesis, writings or doctrine. It does seem to have worked during the cold war. But how sure can we be that it will continue to work for the future?

Deterrence carried a cost. It left behind not just thousands of surplus bombs and tons of fissile materials, but the widespread perception of the nuclear weapon as an icon that is symbolic of national greatness.

We saw this way of thinking explode over India and Pakistan. We saw it glimmer and die in South Africa (for domestic reasons) and Iraq (as the result of international pressure). Now, as a number of states consider the prospects of 'going nuclear', we are attempting to deal with new outbreaks in North Korea and, almost certainly, Iran.

It is important that we do not view nuclear deterrence as an enduring solution, but rather as a practical expedient until we can find a route to permanent peace. It does not address the underlying factors that lead to a crisis in the first place. In addition — as we have seen in the Indian subcontinent — it can have the unfortunate result of encouraging nuclear weapon development in nations that think that they are being 'deterred against'.

I strongly believe that the advantage of nuclear deterrence reached a peak with the end of the cold war more than a decade ago, and has been fading ever since. The United States has been leading an international effort to reduce existing nuclear stockpiles,



E. DUNAND/AP/NEWS.COM

PACH BROTHERS/CORBIS





Can regional alliances, such as NATO, be used to monitor and control the development of nuclear arms?

including its own, and is attempting to stem the tide of nuclear proliferation.

### Realistic expectations

Where do we go from here? The nature of humankind is too complicated, and replete with examples of our inability to organize ourselves as world citizens, for us to put much hope in being able to truly 'outlaw' war and aggression any time soon. Thus, unlike the framers of the Baruch Plan, I have never put much faith in the notion that 'complete and total disarmament' is a realizable goal in the near-term.

Instead, the path to peace probably lies in forging alliances along the lines of the North Atlantic Treaty Organization (NATO). This model could then be extended to form multiple-country alliances in every corner of the globe, with priorities in Southeast Asia, then the rest of Europe and Asia, South America, then the Middle East, and Africa.

All indications point to Southeast Asia — where economic development and technical sophistication is burgeoning — as an area with the potential to become a significant problem in terms of nuclear proliferation. Abdul Qadeer Khan, the architect of Pakistan's nuclear programme, centred his illicit manufacturing operations in Malaysia — they included the production of centrifuges for highly enriched uranium. In addition, recent revelations about activities in North Korea and South Korea indicate that they have long been exploring the development of nuclear weapons. I would suggest that, in this respect, they are not alone among nations in this region.

It is worth noting that all NATO nations, including the most recent additions, are capable of building nuclear weapons, but have not done so because the alliance provides nuclear burden-sharing and protection. With time, the vast majority of the nations of Southeast Asia could acquire or build nuclear weapons, and the Non-Proliferation Treaty (NPT) alone is not sufficient to stop

them from doing so. But a collective security agreement that includes a nuclear alliance could provide a solution. In this region, the South-East Asia Treaty Organization, in which the United States has been and remains a major supporter, represents a forum from which to begin wider security agreements.

One might naturally wonder whether an East Asian 'NATO' should (or could) include North Korea. On this point, drawing from the lessons of history, I am optimistic. Recall that NATO did not originally include all European nations. But given time without war, these nations turned towards democracy and, with the fall of Soviet communism, all applied for membership. A similar pattern could be repeated with regard to North Korea.

As for NATO itself, Russia's recent movements away from certain democratic ideals are major obstacles to its membership. However, these trends are easily reversible. For Russia remaining an ally of the West, rather than risking an adversarial relationship, is essential to a stable future.

### Urgent action needed

At the moment, the Middle East presents the greatest challenge to the goal of stitching the world together and arriving at a plan analogue to that of Baruch. After a long history of state-to-state and domestic conflict, aligning its constituent nation states will present a difficult challenge. Clearly, Israel and its supposed substantial arsenal of nuclear weapons would have to be included in a Middle East security pact, rather than remaining (as seen through Arab eyes) an egregious exception.

The go-it-alone mentality that now pervades the Middle East needs to be replaced with a collective security viewpoint. Time is not on our side. I place Southeast Asia first in my pact-making priorities for the reason

that its technical sophistication, as well as its wealth, make it possible for any of its states to acquire a nuclear weapon in a year or two. It would take somewhat longer for Middle Eastern states to reach that point, but only 'somewhat' longer — years, not decades. Let no one doubt that there is real urgency here.

Trying to create a satisfactory non-proliferation regime in the near-term by fixing flaws in the NPT, although commendable, will probably fall short of what is needed. The treaty suffers basic structural problems, as it freezes inequities by naming just five states as legal possessors of nuclear weapons, but all others are not. The serious problem of what to do with the undefined situation for India, Pakistan and Israel — which are not recognized as nuclear weapon states in the treaty — needs to be solved, and soon. The treaty also suffers because it has insufficient checks on cheating, which has led to the emergence of major loopholes that must be closed.

### Looking for leadership

In short, I believe the world is not yet ready to address and fix all the difficulties of the NPT, and that it would be naive to try to expand the NPT into a Baruch Plan regime at this time. Instead, we should concentrate on regional alliances, continent by continent, until a worldview of collective security has emerged.

The IAEA currently serves as a forum for debate on nuclear weapons issues for the nuclear haves and have-nots. My view is that everybody could become a 'have' by taking part in alliances based, more or less, on the NATO model. Intrusive nuclear inspection will become much more feasible once a global network of collective security agreements is in place.

This practice could result in a considerable reduction of fears concerning security. We could then begin to reformulate the NPT and strengthen the IAEA, so that it becomes an effective nuclear policing entity with worldwide jurisdiction.

The United States originally sought to implement the Baruch Plan through the United Nations, but unfortunately this is not now an organization that can satisfactorily demonstrate security leadership. In addition, there is no reliable mechanism through which UN resolutions can be effectively enforced through military means. Hopefully, international institutions will, with time, gain the level of competency needed to earn the level of trust to allow such a proposal to be seriously considered. In the meantime, there is much we can do to achieve the ideals of the Baruch Plan by incremental means. ■

C. Paul Robinson is director and president of Sandia National Laboratories, Albuquerque, New Mexico 87185-0101, USA.

# When greens see red

An environmental warning and call to curb consumerism.

## Red Sky At Morning: America and the Crisis of the Global Environment

by James Gustave Speth

Yale University Press: 2004. 304 pp.

\$24, £15.99

Dick Taverne

James Speth is not someone whose warnings that our planet's environment is rapidly deteriorating can be lightly dismissed. He is a former administrator of the United Nations Development Programme, was an adviser on environmental matters to US presidents Jimmy Carter and Bill Clinton, and was a founder of the World Resources Institute, a valuable source of information about the environment. He is not an eco-fundamentalist blind to any developments that do not suit his case.

In *Red Sky at Morning* he concedes, for example, that in countries of the Organisation for Economic Co-operation and Development, the discharge of chemical waste and sewage into waterways, human exposure to lead, and emissions of sulphur dioxides and particulates have all declined. What's more, he admits that the ozone layer could recover by the middle of this century and that multinational companies often have a better environmental record than their local counterparts. Indeed, he even says that a move away from carbon-intensive fuels may have to include a shift to nuclear power, both fission and fusion (although in a footnote he rather discounts its feasibility).

Although there are flashes of optimism, they are occasional patches of blue in a darkening sky. On the whole, Speth's theme echoes those of the classical prophets of doom: Paul Ehrlich, Barry Commoner, *The Limits to Growth* by Donella Meadows and others, Lester Brown and the Worldwatch Institute, all of whom he quotes with approval. Most of the mantras of green activists are repeated in the pages of his book: the threat to our health from dioxins and pesticides, DDT and endocrine disrupters, and probably, he implies, from genetically modified crops. He extols the virtues of organic farming, stresses the importance of the precautionary principle, and uncritically assumes that globalization damages our environment and increases poverty and inequality.

Speth lists ten principal concerns, most of which are shared by pragmatic environmentalists: the hole in the ozone layer; climate change; the loss of agricultural land; the depletion of tropical forests; the mass extinction of species; population growth; mismanagement of freshwater resources; overfishing



A false picture? Richard Wilson's painting conjures up a rural idyll, but life was never really like this.

and marine pollution; mismanagement of pesticides and organic pollutants; and acid rain and other atmospheric pollutants. Unfortunately, the reaction of governments has been a problem-defined approach, he argues, whereas the underlying causes of the deterioration are consumerism and economic growth. Our fundamental error, he says, is to take an anthropocentric view of the world, to think that "nature belongs to us, not we to nature". Some of Speth's proposed remedies are sensible, but others include unquestioning support for green non-governmental organizations and for the World Social Forum at Porto Alegre in Brazil, the main theme of which is unqualified opposition to capitalism. "It is time," he declares, "for we, the people, as citizens and as consumers, to take charge."

It is perhaps surprising to find a man with Speth's record resurrecting the doctrine of the doomsters of the 1970s: that we will soon exhaust Earth's limited resources. Such forecasts have proved wildly inaccurate. *The Limits to Growth* predicted that we would run out of gold, zinc, mercury and oil before 1992. In 1981, Ehrlich forecast that half of the world's species would be extinct by 2000 and all would be gone by 2010–25. He also wrote, in *The Population Bomb* in 1968: "The battle to feed humanity is over. In the course of the 1970s the world will experience starvation of tragic proportions — hundreds of millions of people will starve to death."

Remedies prescribed by doctors who continually misdiagnose diseases should not

be swallowed uncritically. Speth shows as little regard for contemporary evidence as he does for the reliability of previous forecasts of doom. To promote organic farming as he suggests, using a system with much lower crop yields than other farming methods, would aggravate the shortage of good farming land. He repeats common misconceptions about the harmful effects of the pesticide DDT on human health, which have never been proved. He ignores the fact that, according to the World Health Organization, DDT has prevented more than 50 million deaths from malaria, and that the ban on its use in most countries, even for spraying the inside walls of houses, still results in millions of avoidable deaths.

Leaving aside oversimplified and exaggerated assumptions about the harm to humans caused by chemicals such as dioxins, endocrine disrupters and pesticides, the basic argument of *Red Sky at Morning* is that humans are destroying nature's capital — hence its obsession with limits. This reflects a widely held view in environmentalist circles that the Enlightenment was a disaster because, following Bacon and the birth of modern science, we have sought to control nature, causing its destruction. Speth seems to assume that the pre-Enlightenment era was a kind of blissful Arcadia in which happy peasants lived in harmony with nature, since destroyed by science and technology. In fact it was a time when life was nasty, brutish and short, not least because medieval, pre-scientific superstition could not control cholera,



pestilence and the ravages of any number of diseases that modern medicine, by controlling nature, can prevent or treat. In pre-scientific agrarian society, life was a perpetual struggle in which the scarcity of resources led to a hierarchical, authoritarian form of organization in which most people suffered or prospered according to prescribed rank.

Indeed it is precisely because humanity has learnt to control some aspects of nature that civilization has advanced. Economic growth is the child of science and technology. To cry "Enough is enough", as Speth would have us do, would be to abandon any hope of substantially diminishing poverty, hunger and disease in the world. Without the contribution of science and technology, the task of preserving and enhancing the quality of life on our planet would be impossible. Speth raises serious issues, but they deserve a more balanced treatment than the prescriptions in his book. ■

Dick Taverne is chairman of the charity *Sense About Science*, 60 Cambridge Street, London SW1V 4QQ, UK.

## Crossing the boundary

### Biological Physics: Energy, Information, Life

by Philip Nelson

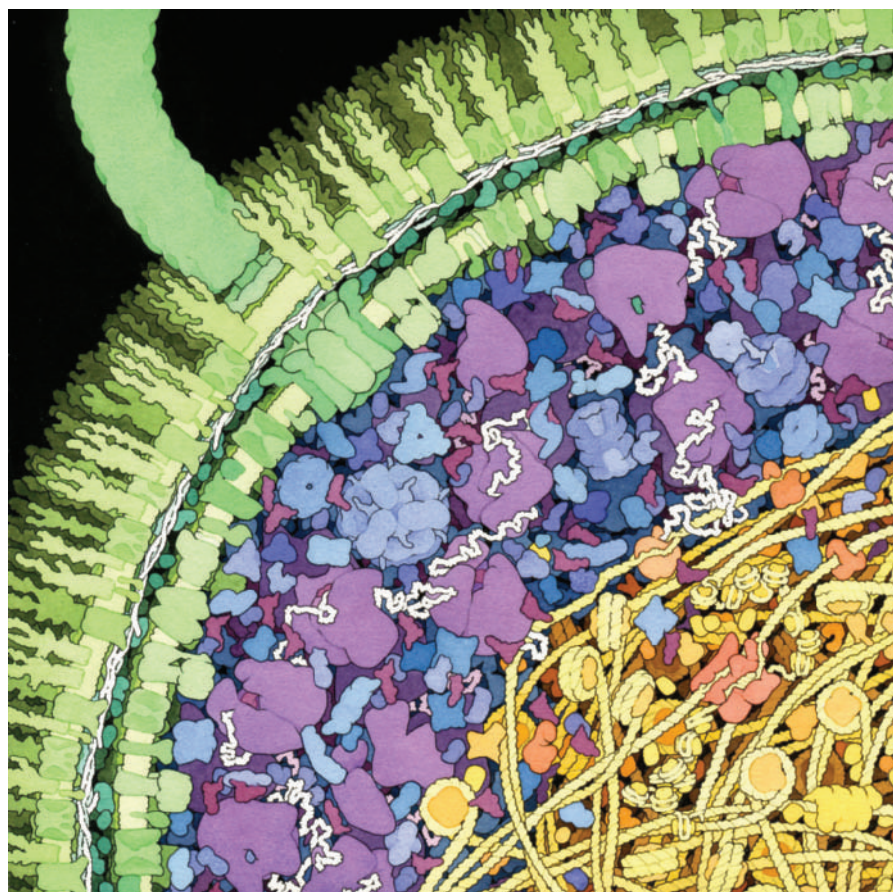
W. H. Freeman: 2004. 600 pp. £34.99, \$106.50

Christopher M. Dobson

"Physics is what physicists do." This definition, which is current in at least some physics departments, is not intended (we trust) simply to confuse the rest of the scientific community. Rather, it suggests that the discipline of physics, more conventionally defined (in the *Shorter Oxford English Dictionary*) as the branch of science "that deals with the nature and properties of matter and energy", can be brought to bear on a wide range of problems across the broad spectrum of human knowledge.

Indeed, 100 years ago, physicists brought about a revolution in chemistry, when the discovery of the electron led to an understanding of the nature of atoms and of molecular structure and bonding. And 50 years ago, physicists ushered in a revolution in biology, with dramatic demonstrations that X-ray diffraction could reveal the structures of DNA and proteins. Nevertheless, relatively few physicists have been "doing biology", and many of these migrated from mainstream physics departments to set up enterprises such as the Laboratory of Molecular Biology in Cambridge, UK, or the Institute of Protein Research in Puschino, Russia.

But times are changing. Physicists in physics departments are now increasingly



Living colour: David Goodsell's 'crowded cell' image of the bacterium *Escherichia coli*.

looking around for grand challenges to rival those that led to major advances in our understanding of many aspects of the physical world, from the fundamental particles of matter to the origins of the Universe. And biology, the study of the nature of life itself, is not short of such challenging problems. Some of these are likely to succumb to bigger and better physical techniques; others might well be revolutionized by the reductionist approach so favoured by physicists to identify and tease out fundamental principles from the vast arrays of data that such techniques can generate.

If such grand challenges are to be tackled successfully, however, it is crucial that physicists collaborate with biologists, and with those in other disciplines, including chemistry, mathematics and medicine. For such collaborations to work, there has to be a common language and an appreciation of the strengths and weaknesses of each other's abilities.

Philip Nelson's excellent text, *Biological Physics*, provides material for the types of course we should now be offering to all our students. Nelson says that his book is for "life science students who are willing to use calculus" and for "physical science and engineering students who are willing to think about cells". And indeed, his book should educate and intrigue both groups. Particularly impressive is the subtle way that topics that often cause

the eyes of even the most diligent student to glaze over are made to seem not just interesting to read about but compelling to learn. Statistical mechanics and thermodynamics are obvious examples, and key concepts such as the Boltzmann distribution and free energy are introduced to explain such questions as why "cells do not work best at the coldest temperatures", despite the fact that life is about generating highly ordered structures. Moreover, although much emphasis is placed on understanding ideas, the text does not allow anyone to forget that some real equations need to be mastered if such an understanding is to be satisfying and widely applicable.

Another good aspect of this book is that it is right up to date in the topics and examples that it covers. Many relatively recent ideas, such as the use of free-energy surfaces for understanding protein folding and of optical tweezers for probing muscle action, are covered in depth, with at least some references to the original literature. There are frequent comments about the way that modern ideas and techniques have emerged. As Nelson cleverly shows, many of the most important scientific breakthroughs have come about unexpectedly through just the type of interdisciplinary studies that this book promotes.

For example, it was the renowned physicist Robert Hooke who realized in the mid-seventeenth century, after peering down a microscope that he himself had constructed,



that living systems are made up of discrete compartments that he called cells. And looking down another microscope in the early nineteenth century, botanist Robert Brown saw pollen grains that were “evidently in motion”. When further experiments showed that the movement was not the result of some sort of life force, he realized that he was observing a general property of finely divided matter. His observations of ‘brownian motion’ — the result of collisions between dynamic particles — were later to be interpreted by Einstein in his doctoral thesis,

work that laid the foundations for quantum theory. It is compelling stuff.

So is this book a perfect work of science? Perhaps almost so, although at first sight a biologist might not think so. Nelson’s book makes frequent use of drawings of biological structures, notably many by David Goodsell, including his famous image of a ‘crowded cell’. By contrast, most modern biology texts, even relatively inexpensive ones, have been transformed by the power of modern microscopy and sophisticated molecular-graphics programs.

Still, the handful of colour pictures in this book demonstrate that one of the many attractions of living systems is frequently their great beauty. And simple observations of biological systems have long influenced the progress of physics: after all, the moving object said to have triggered Newton’s understanding of gravity was undoubtedly of biological origin. ■

Christopher M. Dobson is at the Cavendish Laboratory and the Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK.

## Theatre

# A stage of evolution

## The Darwin Variations

by Jean-François Peyret & Alain Prochiantz  
Performed in French at the Théâtre National de Chaillot, Paris, until 18 December 2004.

Laura Spinney

Darwin’s theory of natural selection has been handed down to us as the broad vision of a brilliant man, meticulously researched, comprehensive and self-assured. We hear less about his doubts, his procrastination and anguished discussions with contemporaries. Nor is it well-known that, in confronting the implications of his theory, he made himself ill. “I’m sick to the stomach,” he confesses at times throughout *The Darwin Variations*.

This play sets the science in the context of the man and his time. The process of publication tends to strip away all the emotion, poetry and music of a discovery. So Alain Prochiantz, a developmental neurobiologist at the Ecole Normale Supérieure, Paris, and theatre director Jean-François Peyret have set out to recreate it. They have given Darwin the poetic licence that Ovid took for granted when he wrote his poem *Metamorphoses*. Like Ovid, the actors in *The Darwin Variations* are free to imagine the future transformation of humanity — even a humanity that takes charge of its own evolution. Their musings intertwine with Darwin’s in a dialogue that takes in old and new ideas, about where to draw the line between man and animal, and about consciousness, genetic engineering and interfaces between brain and machine. Over all of this hangs the threat of the inevitable conclusion: that there is no grand design, no notion of perfection to which evolution is propelling us.

This conclusion is embodied by the action on stage. This is theatre of the absurd, where nothing happens for a reason. Actors move about in a disconnected way, apparently oblivious to an off-stage dialogue. The speakers are invisible, but their voices seem to press in on the set, clamouring to be let in. Every so often they succeed, and the actors respond.



*The Darwin Variations*, featuring Irene Jacob (right), ponders the future transformation of humanity.

The curtain rises, for instance, on an empty stage. Two actors appear, wearing dark glasses, carrying a bench, a pebble and an egg. Off stage a blind man on his death-bed debates the existence of God with a priest, based on their different experiences of the natural world. The conversation shifts, and the question becomes whether, given a choice, a blind man who knows the world by touch alone would choose to have eyes or a longer pair of arms. Just then, one of the blind men on stage throws up the egg, and the other catches it in mid-air.

In another scene, a woman sitting on a bench, cutting up newspaper, asks the man beside her if contempt is expressed by a slight pouting and a flaring of the nostrils with a small expiration — one of Darwin’s observations from his book *The Expression of the Emotions in Man and Animals*. The man doesn’t respond. She asks him, indignantly, if the expression for disgust resembles that

of someone about to spit. He gazes off into the distance. Enraged, she threatens him with her scissors. His face contorts, he throws himself at her, and the scene ends in a fight to the death.

The play is funny, original and anything but didactic. It takes familiar ideas and pursues them through their serious implications *ad absurdum*. The third in a trilogy based on metamorphoses, *The Darwin Variations* is, in a sense, a product of evolution itself. The actors began with a script and moulded it into a piece of theatre, according to their own continuously evolving ideas. It will probably continue to evolve through its run at the Théâtre National de Chaillot in Paris. Those attending on the last night may see something quite different to those present at the opening — not necessarily any better or worse, just different. ■

Laura Spinney is a freelance writer based in London and Paris.

# Crazy, but correct

How a non-conformist theory beat scepticism and got into the textbooks.

**Daniel E. Koshland Jr**

**T**he event that marked the biggest change in my career was the postulation of a controversial theory at a fairly early stage of my working life. In the late 1950s, I was working as an assistant professor at Brookhaven National Laboratory, New York. I was studying the muscle enzyme hexokinase, and was puzzled as to why water — an analogue of glucose, the enzyme's natural substrate, and present at a much higher concentration — did not react with hexokinase. I began to think about the role of water in general and realized that the major obstacle of a living system was not so much the need to 'activate' water to react when needed, as the need to prevent water from reacting in cases where it was not needed to do so. It then dawned on me that any mechanism for hexokinase activity based on Emil Fischer's long-established 'lock and key' theory (sometimes called the 'key-lock' or 'template' theory) for enzyme-substrate interactions would result in catastrophic losses to side reactions in real life. Moreover, the logic behind this realization about water would also apply to other cases where a smaller analogue of the true substrate failed to react with an enzyme.

I went back over my thoughts and experiments. Finding no errors, I decided that the template theory needed revision. Instead of the wooden jigsaw-puzzle analogy of fitting for the template model, my concept was more like the fit of a hand in a glove, a moderately flexible enzyme (the glove) fitting a moderately flexible substrate (the hand). I felt this was the better analogy for proteins and substrates. Speaking very simplistically, in the 'induced-fit' model, water could not substitute for the larger glucose to induce a conformational change in the enzyme, allowing it to react. Whereas in the template model, all the catalytic groups are already aligned so that water could react. In addition, my theory explained so many previously unexplained phenomena that I felt it had to be right.

Realizing that a young scientist was largely identified as promising on the basis of one or two good papers — and could be largely destroyed on the basis of one or two bad papers — I was aware of the risks I was taking. But I had become so enamoured of what I called the 'induced-fit' theory that I decided to publish it.

I remember giving a talk about it at a meeting of the American Chemical Society. As I walked out of the hall, I heard a couple of young scientists say, "Koshland used to do



Daniel Koshland (above) took a gamble when he challenged the theories of Emil Fischer (right).

good work, but he seems to have gotten senile prematurely." After an invited lecture in Gatlinburg, Tennessee, the applause was polite, but tepid. I asked fellow enzymologist, Ron Breslow, what he thought of the audience reaction. He said, "Dan, if you should turn out to be right, at least you can say that no one in that room believed you at the time."

Some journals were skeptical and rejected my manuscripts with statements such as, "The theory of Emil Fischer has been a cornerstone for 100 years and will not be overturned by the ideas of an obscure young biochemist from a young national laboratory." However, other journals did agree to publish my work, so the theory became part of scientific discourse and I received many invitations to talk about it.

The theory also explained many other enzymatic phenomena — such as hormone action, feedback inhibition, cooperativity and receptor function — so its acceptance grew slowly, but steadily. Its progress was accelerated a few years later when X-ray crystallography (initially carried out in the laboratories of Thomas Steitz and William Lipscomb) revealed pictures of enzyme-substrate interactions in carboxypeptidase and hexokinase occurring exactly as predicted by the induced-fit theory. The success of the theory emboldened me, an organic chemist, to venture into new biological areas, such as sensory signalling, which I would not have dared to enter before.

Now, many years later when the theory is explained in textbooks of biochemistry, it is enlightening to look back and learn from the progression of a theory from unconventional hypothesis to conventional wisdom.

I remember my angry reaction and frustration at the journals that rejected my articles. I realize now that a new theory is likely to meet resistance, but it should, if based on good experiments, receive sceptical encouragement if science is to remain in balance. Non-conformists are necessary for progress in science, just as mutations are necessary for progress in evolution. However, there must be constraints to select good mutations from bad mutations.

Too many mutations block evolution, as error-prone strains of bacteria have proved.

So non-conformist thinking in science must be encouraged to make progress, but restrained to prevent anarchy. In science, it is peer-reviewed journals and granting agencies that provide such balance.

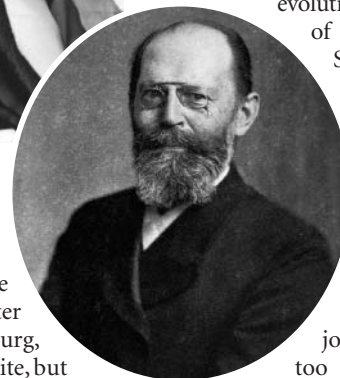
The trouble is that journals can easily become too conservative, because

editors find it easier to reject the unusual than to take a chance on the unthinkable. Later in life, when I became editor-in-chief of the journal *Science*, my early experience allowed me to keep a friendly eye out for the non-conformist. But it is not easy to select between the unexpected and the impossible in today's world of increasing specialization and exponential increase in knowledge. The existence of multiple journals provides the final safeguard against too much conservatism and is the ultimate reason that science is more receptive to non-conformity than any other segment of our society. But does science have any lessons for non-conformism in other spheres, such as politics and religion?

Non-conformity is looked on with more hostility by religion, government and culture than by science — because each of them is more vulnerable to change than science is. The other segments of our society have yet to find a better mechanism for encouraging non-conformity to achieve progress, while still controlling non-conformity to prevent chaos. Science has achieved the best balance, but it must fight to preserve this and serve as a beacon to other sectors of our society. ■

Daniel E. Koshland is in the Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA.

CORBIS





# Don't lose your reputation

Ernst Fehr

Collective action in large groups whose members are genetically unrelated is a distinguishing feature of the human species. Individual reputations may be a key to a satisfactory evolutionary explanation.

When the Allied forces invaded Normandy during the Second World War, thousands of people were involved in the preparations and in the invasion itself; a similar number of Germans were probably involved in defending the occupied territory. War is a prime example of large-scale within-group cooperation between genetically unrelated individuals (Fig. 1). War also illustrates the fact that within-group cooperation often serves the purpose of between-group aggression. Modern states are able to enforce cooperation in large groups by means of sophisticated institutions that punish individuals who refuse to meet their duties and reward those who follow their superiors' commands. The existence of such cooperation-enhancing institutions is very puzzling from an evolutionary viewpoint, however, because no other species seems to have succeeded in establishing large-scale cooperation among genetically unrelated strangers<sup>1</sup>.

The puzzle behind this cooperation can be summarized as follows. Institutions that enhance within-group cooperation typically benefit all group members. The effect of a single group member on the institution's success is negligible, but the contribution cost is not negligible for the individual. Why, therefore, should a self-interested individual pay the cost of sustaining cooperative institutions? More generally, why should a self-interested individual contribute anything to a public good that — once it exists — an individual can consume regardless of whether he contributed or not? On page 499 of this issue<sup>2</sup>, Panchanathan and Boyd substantially advance the scope of reputation-based models<sup>3–5</sup> and show that individuals' concern for their reputation may be a solution to this puzzle.

Evolutionary psychologists have sought to answer the puzzle of human collective action for decades. However, progress was limited because of a lack of commitment to mathematically rigorous theorizing. Many researchers erroneously thought that Trivers's notion of reciprocal altruism<sup>6</sup>, which Axelrod and Hamilton successfully formalized as a tit-for-tat strategy for two-person interactions, provides the solution to the problem. Trivers himself speculated that reciprocal altruism "may favour a multiparty altruistic system in which altruistic acts are dispensed freely among more than two



Figure 1 Call to arms. Why do humans cooperate with others who are not genetically related to them, particularly in large-scale activities such as the waging of war? Panchanathan and Boyd<sup>2</sup> suggest that each individual is motivated by the desire to maintain their reputation as a contributor to the public good.

individuals". However, it is always easier to speculate than to provide a rigorous model, and the speculation is likely to be wrong in this case.

In the context of the problem of public-goods provision, a reciprocally altruistic individual is willing to contribute to the public good if sufficient numbers of other group members are also willing to contribute. Unfortunately, the presence of only a small number of defectors quickly causes cooperation to unravel if it is solely based on conditionally cooperative behaviour, because the defectors induce the conditional cooperators to defect as well. Theory and simulations suggest that reciprocally altruistic strategies can only sustain high levels of cooperation in two-person interactions<sup>7</sup>. Moreover, experimental evidence indicates that cooperation in public-good games typically unravels because it is not possible to discipline 'free riders' — those who take advantage of others' cooperation — if only conditionally cooperative strategies are available<sup>8</sup>.

In contrast to reciprocal altruism, the notion of altruistic punishment has been more successful in explaining collective action, because direct punishment disciplines free riders<sup>8</sup>. Altruistic punishers contribute to collective actions and are willing to sanction individual defectors even if they incur net costs by doing so. However, within-group selection in the presence of altruistic punishers favours cooperative individuals who do not punish defectors. Such individuals will never be punished — because they contribute to the collective action — but they also never bear the cost of punishing defectors. These pure cooperators are thus 'second-order' free riders because they do not contribute to the disciplining of the selfish individuals. Therefore, pure cooperators will crowd out altruistic punishers unless there is group competition that renders groups with a higher share of altruistic punishers more successful<sup>9</sup>.

Panchanathan and Boyd's contribution<sup>2</sup> solves this second-order free-rider problem

CHARLES & JOSETTE LENARS/CORBIS

by linking the notion of indirect reciprocity<sup>3,4</sup> with an individual's reputation for contributing to collective actions. An indirectly reciprocal individual helps another individual if the recipient of the help has a good reputation and if — by helping — the individual can himself acquire or maintain a good reputation. If helping accords a good reputation and individuals with a good reputation will also receive help when needed, the act of helping becomes a self-interested choice. Note that no direct reciprocal interactions are necessary in this case; the only prerequisites are that helping confers a good reputation and that people with a good reputation are helped in the future. In a pioneering experiment, Milinski *et al.*<sup>10</sup> showed that if potential donors in a game of indirect reciprocity are informed of a recipient's contribution in a previously played collective-action game, cooperative behaviour in the collective-action game is sustained at very high levels. Apparently, potential donors do not help those who fail to contribute to the public good but they assist those who contribute. This pattern of helping provides strong incentives for selfish individuals to contribute to the public good.

Inspired by this experiment, Panchanathan and Boyd provide a parsimonious evolutionary model by linking an indirect-reciprocity game with a game of collective action. The key element in their model is the shunning strategy. A shunner always helps a deserving recipient; if he does not, he loses his good reputation. However, the shunner can maintain his good reputation by refusing to help an undeserving recipient. A recipient deserves help if he is in good standing. In contrast, a recipient does not deserve help if he is in bad standing; that is, if he either did not contribute to the collective action or did not help a deserving recipient in previous interactions in the indirect-reciprocity game. Therefore, the shunners punish free riders who did not participate in the collective action without any cost to themselves because the shunners refuse to help free riders when they are in need. In fact, because the shunners save the cost of helping by their refusal to help, this form of punishment is in their self-interest. Thus, if a system of social norms based on the shunning strategy prevails, shunners face no selection pressures — and the second-order free-rider problem is solved.

A crucial element of the shunning strategy is that it rests on the recipients' behaviour in the collective-action game. Panchanathan and Boyd<sup>2</sup> show, however, that a shunning strategy cannot establish itself in a group where the helping decision is not linked to the cooperative decision in the collective-action game. Thus, a system of social norms that does not punish free riders by refusing to help them is just as stable as a system of norms that punishes free riders. A convincing

evolutionary solution to the second-order free-rider problem therefore requires additional mechanisms. One such mechanism could be competition between groups with different social norms, because groups that successfully link the helping decision with individuals' behaviour in the collective-action game are better able to solve their collective-action problems. Group competition therefore does not serve as a mechanism for offsetting within-group selection pressures on shunners — because shunning is an individually advantageous strategy — but is merely a device for the selection of cooperation-enhancing social norms. ■

## Materials science

# A 'bed of nails' on silicon

Max G. Lagally and Robert H. Blick

The future of electronics may rest on devices that integrate other semiconductors with silicon. A means of creating tiny semiconductor pillars on a silicon surface is now demonstrated.

Computers rely on silicon. Although other semiconductors have desirable features, in this context the materials properties of silicon are so outstanding that it is really the only choice for the large-scale integration of fast electronic devices. But there is a dark shadow over silicon, in that it produces no light. Many devices use light — simple ones, such as the solid-state diode lasers typically found in CD players, for example; and more complex ones, such as the light amplifiers used in long-range optical-fibre communication. Gallium arsenide or other more exotic semiconductors must be incorporated into these devices to generate light. In *Nano Letters*, Mårtensson *et al.*<sup>1</sup> present a new means of doing so.

Why is the lack of optical activity in silicon a problem? After all, CD players don't need much computational power. And computers don't need light. Or do they? The ever-decreasing size of transistors made of silicon means that the transit time for charge carriers through them (and hence their switching time) is increasing rapidly. The overall speed of a microprocessor will be more and more limited by the time delay inherent in the connections between individual transistors<sup>2</sup>. According to the International Technology Roadmap for Semiconductors<sup>3</sup>, the traditional copper/dielectric-material system for interconnects will have to be replaced by some novel on-chip interconnect scheme beyond the year 2010.

Optical communication is very fast, so if communication between the far reaches of a chip were possible by optical means, the full advantage of size scaling could be realized. Hence there is a demand for faster on-chip data communication using opto-electronic

Ernst Fehr is at the Institute for Empirical Research in Economics, University of Zurich, Blumliplatzstrasse 10, CH-8006 Zurich, Switzerland. e-mail: efehr@iew.unizh.ch

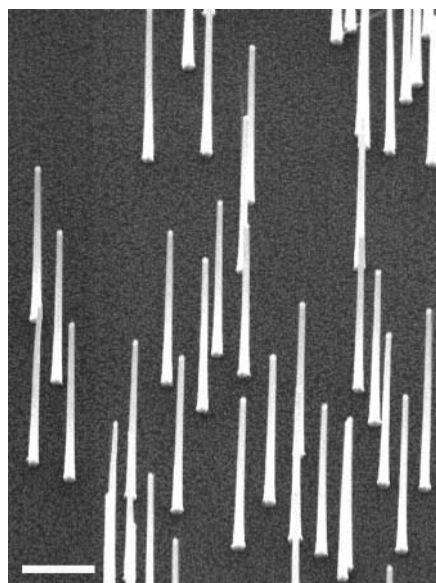
1. Fehr, E. & Fischbacher, U. *Nature* **425**, 785–791 (2003).
2. Panchanathan, K. & Boyd, R. *Nature* **432**, 499–502 (2004).
3. Nowak, M. A. & Sigmund, K. *Nature* **393**, 573–577 (1998).
4. Leimar, O. & Hammerstein, P. *Proc. R. Soc. Lond. B* **268**, 745–753 (2001).
5. Gintis, H., Smith, E. A. & Bowles, S. *J. Theor. Biol.* **213**, 103–119 (2001).
6. Trivers, R. L. *Q. Rev. Biol.* **46**, 35–57 (1971).
7. Boyd, R. & Richerson, P. J. *J. Theor. Biol.* **132**, 337–356 (1988).
8. Fehr, E. & Gächter, S. *Nature* **415**, 137–140 (2002).
9. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535 (2003).
10. Milinski, M., Semmann, D. & Krambeck, H. J. *Nature* **415**, 424–426 (2002).

components, such as light-emitting diodes made from aluminium, gallium, arsenic and phosphorus (elements from groups III and V of the periodic table that are commonly paired in semiconducting 'III–V' materials). These materials are optically active and can be grown as heterostructures (of different materials that are crystallographically linked) to form quantum wells and quantum boxes that emit coherent light through electrical stimulation. Their optical activity can be designed, as can the crystal growth sequence required to obtain it.

Because the fabrication of silicon devices and the process of heterostructure growth for lasers are well developed, devices that require both are typically manufactured using a hybrid technology — one that contains different chip sets of silicon and III–V circuits connected by wires. But it is easy to see that this approach fails to speed on-chip communication. For that, direct crystallographic integration of III–V heterostructures onto silicon is needed, preferably at the nanoscale.

Exactly this approach has now been demonstrated by Mårtensson *et al.*<sup>1</sup>, who have grown optically active III–V nanowire heterostructures on silicon. The nanowires — which sometimes look like a bed of nails (Fig. 1) — are typically 2 µm long, with a base diameter of 50–100 nm. They have light-emitting sections grown into them and thus function effectively as light towers on the silicon substrate. These nanopillars, as well as being a beacon of hope for on-chip optical communication, open a variety of other communication applications by linking optical components with silicon-based circuitry.





**Figure 1** To the point. Mårtensson *et al.*<sup>1</sup> have grown these gallium phosphide nanowires, seeded by gold nanoparticles, on a silicon substrate. The integration of 'III-V' materials such as this with silicon could prove vital to the future of electronics. Scale bar, 500 nm.

Mårtensson and colleagues have grown III-V nanowires on standard silicon wafers by using a 40-year-old technique — vapour-liquid-solid (VLS) growth — and some new tricks. VLS growth works on the principle of using two immiscible materials that have a deep eutectic point in their phase diagram<sup>4</sup> (the eutectic point marks the lowest freezing point of all possible mixtures of the two components). For example, a metal particle on a surface and surrounded by vapour of the other component (gallium phosphide, GaP, in this work<sup>1</sup>) will absorb that vapour until its composition changes enough to become liquid. Because of the continued supply of vapour, solid GaP will precipitate into a column whose diameter matches that of the particle. The particle will always remain at the apex of the column. The new aspects here<sup>1</sup> are the use of sufficiently small nanoparticles, the surface preparation and the growth conditions for the III-V nanowires. Mårtensson *et al.* have made full use of the power of epitaxial growth techniques to construct nanopillars that contain segments of GaAs<sub>x</sub>P<sub>1-x</sub> for light emission.

These III-V materials do not grow epitaxially as films on silicon: the mismatch between their lattice structures soon causes the growth to become three-dimensional (a feature recognized about 15 years ago and developed to provide strain-engineered quantum dots<sup>5</sup>). The authors realized that, if they could make the columnar structure small enough, it could relax laterally to accommodate the strain caused by lattice mismatch and remain epitaxial. They were able to show that the transition from the silicon substrate to the GaP column occurs

without any defects, enabling the vertical growth of the wires to continue.

The luminescence of the GaAs<sub>x</sub>P<sub>1-x</sub> nanopillars is as bright at room temperature as it is at low temperatures, which makes them strong candidates indeed for many applications. This group has already produced other devices<sup>6,7</sup>, but the III-V combination with silicon is of greatest significance. Now conventional silicon transistors with light-emitting components can be realized, perhaps with III-V nanopillars towering over conventional CMOS circuitry and exchanging information. One can also envisage inter-chip communication lines, once surface-emitting lasers have been integrated into the nanopillars.

There are also implications for quantum communication, in which single-electron devices fabricated in nanowires<sup>8</sup> need to exchange information either through single photons or through vibrational modes in nanoelectromechanical devices<sup>9</sup>. One can imagine the creation of 'excitons' (electron-

hole pairs) in the nanopillars that could be entangled through the emission of photons. A pair of nanopillars would thus function as either an emitter or a receiver of entangled quantum-information states. There are many obstacles to achieving this, but the experiments by Mårtensson *et al.*<sup>1</sup> inspire hope for the future. ■

Max G. Lagally is in the Department of Materials Science and Engineering, and Robert H. Blick is in the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.

e-mails: lagally@engr.wisc.edu  
blick@engr.wisc.edu

1. Mårtensson, T. *et al.* *Nano Lett.* **4**, 1987–1990 (2004).
2. Peercy, P. S. *Nature* **406**, 1023–1026 (2000).
3. <http://public.itrs.net>
4. Wagner, R. S. & Ellis, W. C. *Appl. Phys. Lett.* **4**, 89–90 (1964).
5. Woll, A. R., Rugheimer, P. & Lagally, M. G. *Int. J. High-Speed Electron. Syst.* **12**, 45–78 (2002).
6. Björk, M. T. *et al.* *Appl. Phys. Lett.* **81**, 4458–4460 (2002).
7. Dick, K. A. *et al.* *Nature Mater.* **3**, 380–384 (2004).
8. Thelander, C. *et al.* *Appl. Phys. Lett.* **83**, 2052–2054 (2003).
9. Scheible, D. V. & Blick, R. H. *Appl. Phys. Lett.* **84**, 4632–4634 (2004).

## Regenerative medicine

# Prometheus unbound

Michael D. Schneider

The discovery of a protein that stimulates cell migration and survival in damaged mouse hearts suggests a potential new approach to the treatment of heart attacks.

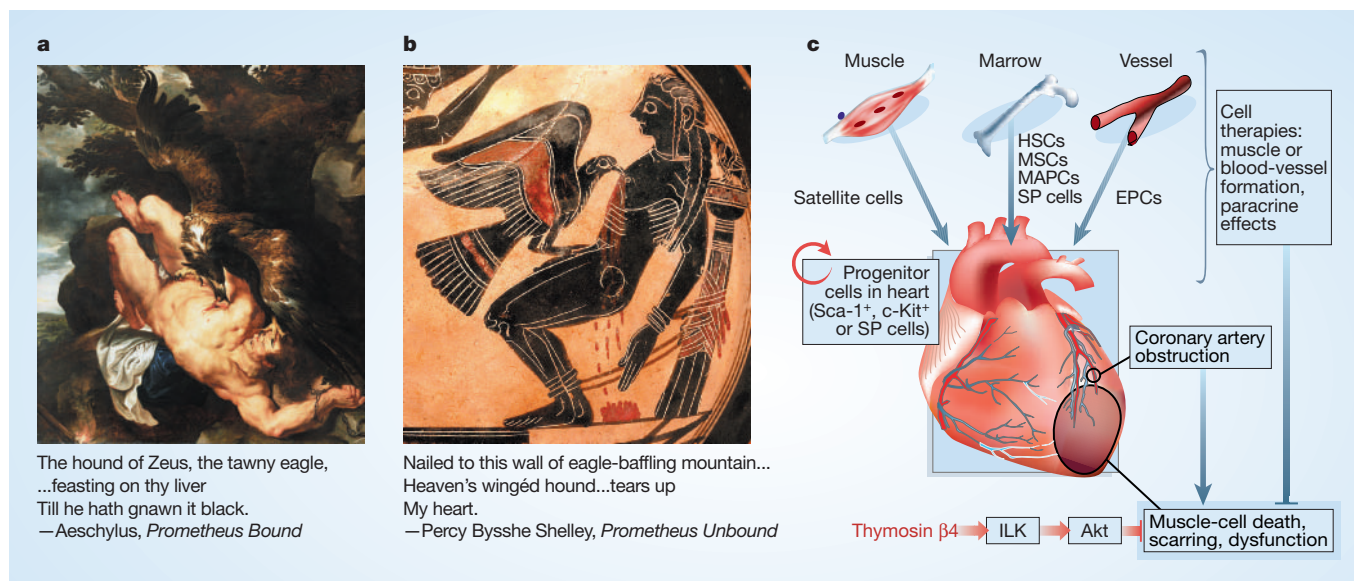
**T**he death of heart muscle cells, unmatched by the generation of replacement cells, is ultimately a basis for death from heart disease. This concept of heart failure as a muscle-cell-deficiency disorder has prompted vigorous investigation of ways to replace or protect the endangered cell, including gene and cell therapies. Writing on page 466 of this issue, Bock-Marquette and colleagues<sup>1</sup> suggest a variation on this theme. They show that thymosin  $\beta_4$ , a protein that binds to components of the cellular internal 'skeleton' and activates a cell-survival signalling pathway, curbs the size of heart attacks in mice. The report provides tantalizing clues towards a workable remedy for this prevailing cause of heart failure.

The legend of Prometheus — the Greek Titan punished by Zeus for the hubris of bringing fire to mankind — provides an icon for regenerative medicine<sup>2</sup>. Shackled to Mount Caucasus by adamantine chains, Prometheus was for 30,000 years feasted upon by an eagle, which devoured his liver daily, only for it to regrow entirely each night (Fig. 1a, overleaf). The recuperative power of the liver, although not quite as impressive as in myth, is nowadays placed on a firm scientific footing<sup>3</sup>, and stands in distinction to the skimpier capacity of heart muscle for self-repair — alternative versions of the legend

notwithstanding (Fig. 1b). The death of heart muscle cells is in large part responsible for the prevalent varieties of heart disease, both acute and chronic<sup>2</sup>.

The Promethean ambition to generate new cells in the heart has recently focused on implanting donor cells of various kinds, in high-profile human trials<sup>4</sup> (Fig. 1c). But the biological foundations for the utility of grafting are still shaky<sup>5,6</sup>. Implanted skeletal muscle cells rarely differentiate into their cardiac cousins, or even couple with them electrically<sup>6</sup>. Meanwhile, haematopoietic (blood-forming) stem cells might give rise to cardiac muscle at most only rarely *in vivo*<sup>7–9</sup>, whereas other components of bone marrow have received less attention. Thus an enigma persists concerning the relative roles of muscle creation, cell fusion, blood-vessel development and the short- or long-range effects of proteins secreted by grafted cells.

A possible portal to better cardiac repair may involve the identification of genes that are active in the developing heart or in 'multipotent' cells (unspecialized cells, with the potential to produce many or most other cell types, that recapitulate the creation of heart muscle and are more accessible to investigation<sup>5</sup>). The gene encoding a secreted protein called thymosin  $\beta_4$  is active in both situations. Thymosin  $\beta_4$  is one of several related



**Figure 1** Cardiac repair in legend, counter-legend, and reality. **a**, The robust ability of the liver to regenerate itself is embodied in the Greek myth of Prometheus. **b**, In alternative versions of the Promethean legend, the heart regrows. In reality, however, the heart shows little capacity for regeneration, and death of heart muscle cells underlies the prevalent forms of heart disease. **c**, Attempts to regenerate cardiac muscle have been made experimentally and in clinical trials, using satellite cells from skeletal muscle; haematopoietic stem cells (HSCs), mesenchymal stem cells (MSCs), multipotent adult progenitor cells (MAPCs) and side population (SP) cells from bone marrow; and endothelial progenitor cells (EPCs) from the bloodstream. Self-renewing progenitor and stem cells from heart muscle — found by stem-cell antigen-1 (Sca-1) or c-Kit expression, or

through SP properties — have an especially strong predisposition to create new heart muscle. Secreted proteins (paracrine effects) might be important, though, to cell-based therapies. In keeping with this, Bock-Marquette *et al.*<sup>1</sup> have found that, after the coronary artery is obstructed in mice, injection of the secreted protein thymosin  $\beta$ 4 reduces muscle-cell death, scarring and heart dysfunction. Thymosin  $\beta$ 4 seems to work by activating a cell-survival pathway, probably involving integrin-linked kinase (ILK) and Akt. (**a**, Painting by Peter Paul Rubens, c.1611–12, by permission of the Philadelphia Museum of Art/Corbis. **b**, Spartan vase attributed to the Painter of Archelais II, c.555 BC, Vatican, Gregorian Museum of Etruscan Art, by permission of Photo Scala, Florence.)

proteins, first purified from the thymus, that bind to cytoskeletal filaments containing the actin protein. Specifically, it sequesters actin monomers and stops them from assembling onto the ends of polymeric actin filaments<sup>10</sup>. Its actin-binding motif is present in nearly 40 proteins that modulate actin-based cell motility, and thymosin  $\beta$ 4 itself promotes the migration of endothelial cells (which make up the lining of blood vessels) and skin cells. It is this property that underlies its auspicious ability to promote blood-vessel formation and skin wound healing<sup>11</sup>. Improved healing requires no more than the central seven-amino-acid actin-binding motif<sup>11</sup>.

The enriched expression of thymosin  $\beta$ 4 in the developing heart, and its wound-healing properties, incited the interest of Bock-Marquette and colleagues<sup>1</sup>. They now report that this protein markedly enhances cardiac repair in mice if administered immediately after the coronary artery is obstructed experimentally. They find that thymosin  $\beta$ 4 stimulates the migration and survival of heart cells, working through a known survival signalling pathway<sup>12</sup> (Fig. 1c).

Bock-Marquette *et al.* report that, in embryonic mouse hearts, thymosin  $\beta$ 4 is present chiefly in migratory cells — subsets of muscle cells and endothelial cells — that fashion the cardiac valves, spongy trabeculae and outflow tract. The authors also show that

in culture, thymosin  $\beta$ 4 can bind to, and be internalized by, embryonic heart cells, with pronounced effects on the migration of embryonic endothelial and cardiac muscle cells. Not only that, but the survival of cultured cardiac muscle cells from newborn rats is also markedly enhanced.

A molecular search for proteins that bind thymosin  $\beta$ 4 led the investigators to the cytoskeletal protein PINCH, and from PINCH to its partner, integrin-linked kinase (ILK). Both proteins are essential constituents of the 'focal adhesion complex' that allows cells to interact with their extracellular matrix. Another enzyme, Akt (also known as protein kinase B), is a known substrate of ILK, with wide-ranging signalling functions that affect growth, survival and motility<sup>12</sup>. Bock-Marquette *et al.* found that extracellular thymosin  $\beta$ 4 activates Akt in a manner that seems contingent on ILK and independent of actin dynamics.

Consistent with evidence that increased levels or activity of Akt render the heart resistant to damage by oxygen deprivation<sup>13</sup>, the authors found that administering thymosin  $\beta$ 4 to mice undergoing coronary-artery obstruction reduced scar volume, inhibited ventricular dilatation, and salvaged ventricular pump function. These benefits accrued within three days of injury, with the suggested mechanism being activation of Akt and consequent prevention of the death of adult

cardiac muscle cells, rather than a conceptual alternative — recruitment of adult progenitor or stem cells to, or within, the heart<sup>5,14</sup>. (That might have been a possibility because, at least in regenerating skeletal muscle, stem-cell recruitment is enhanced by insulin-like growth factor-I, one of the best-known inducers of Akt function<sup>15</sup>.)

As for poor Prometheus, his liver and heart were finally saved from their lot by Herakles, who slew the Caucasian eagle. But what will be the fateful arrow in the armamentarium of cardiac repair? Akt has shown early promise both in the form of a virus-encoded gene delivered to the heart<sup>13</sup> and as a genetic enhancement of transplanted mesenchymal stem cells<sup>16</sup>. If proven to be safe (noting that all cells contain monomeric actin), and if extrapolated successfully to mammals larger than mice (a common stumbling-block in bench-to-bedside translation), the activation of endogenous Akt by thymosin  $\beta$ 4, or in principle by equivalent small molecules, might prove more easily workable than variations on this theme requiring exogenous Akt, cells or other genes. Bearing in mind the high likelihood that cellular therapies for cardiac repair act through secreted molecules, it will be intriguing to see whether  $\beta$ -thymosins are not merely sufficient, but might also be necessary. ■

Michael D. Schneider is at the Center for Cardiovascular Development, Departments of



Medicine, Molecular and Cellular Biology, and Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, Texas 77030, USA.  
e-mail: michael@bcm.tmc.edu

1. Bock-Marquette, I. *et al. Nature* **432**, 466–472 (2004).
2. Rosenthal, N. *N. Engl. J. Med.* **349**, 267–274 (2003).
3. Borowiak, M. *et al. Proc. Natl Acad. Sci. USA* **101**, 10608–10613 (2004).
4. Mathur, A. & Martin, J. F. *Lancet* **364**, 183–192 (2004).
5. Olson, E. N. & Schneider, M. D. *Genes Dev.* **17**, 1937–1956 (2003).
6. Leobon, B. *et al. Proc. Natl Acad. Sci. USA* **100**, 7808–7811 (2003).
7. Balsam, L. B. *et al. Nature* **428**, 668–673 (2004).
8. Murry, C. E. *et al. Nature* **428**, 664–668 (2004).
9. Nygren, J. M. *et al. Nature Med.* **10**, 494–501 (2004).
10. Hertzog, M. *et al. Cell* **117**, 611–623 (2004).
11. Philp, D., Goldstein, A. L. & Kleinman, H. K. *Mech. Ageing Dev.* **125**, 113–115 (2004).
12. Brazil, D. P., Park, J. & Hemmings, B. A. *Cell* **111**, 293–303 (2002).
13. Fujio, Y., Nguyen, T., Wencker, D., Kitsis, R. N. & Walsh, K. *Circulation* **101**, 660–667 (2000).
14. Oh, H. *et al. Proc. Natl Acad. Sci. USA* **100**, 12313–12318 (2003).
15. Musaro, A. *et al. Proc. Natl Acad. Sci. USA* **101**, 1206–1210 (2004).
16. Mangi, A. A. *et al. Nature Med.* **9**, 1195–1201 (2003).

## Quantum information

# Atomic recorder for light quanta

Jean-Michel Raimond

The quantum information carried by a faint laser pulse has been trapped in a gas of atoms. This 'quantum memory' paves the way for networks that transmit and process information in non-classical ways.

Most of the information we get through the World Wide Web travels encoded on inch-long laser pulses rushing at light-speed down hair-thin glass fibres many thousands of leagues under the seas. One day, that information might be coded onto the quantum properties of these pulses, the weird rules of quantum logic opening a wealth of new possibilities<sup>1</sup>. Processing this information would require that it be copied from the light onto motionless objects, to be stored. But quantum states are fragile, and copying them is not easy. In a step towards the realization of a quantum-information network, Julsgaard *et al.*<sup>2</sup> demonstrate just such a quantum memory, in which the state of a faint laser pulse is faithfully copied onto an ensemble of atoms (see page 482 of this issue).

The power and strangeness of the quantum arise from a few striking properties. Quantum systems can be in a superposition of states, allowing quantum bits (qubits) to take two logical values at once. Quantum states cannot be cloned — a copy operation inevitably destroys the original. This is a key point for quantum cryptography<sup>3</sup>, as an eavesdropper cannot access quantum information without revealing his presence. Quantum systems can be entangled, forming a single entity whatever the physical distance between them, and these weird correlations are used for teleportation<sup>4</sup> — a quantum 'fax machine' that transmits a quantum state independently of the particle that carries it.

Light quanta (photons) or faint laser pulses are excellent carriers of quantum information. They travel unaffected over long distances, are easily read out in detectors, and hence have been thoroughly exploited for quantum tests<sup>5</sup>, cryptography<sup>3,6</sup> and teleportation<sup>4,7</sup>. For many purposes,

however, their main quality — that they travel at the speed of light — is an inconvenience, as they cannot be stored for any extended time; the best optical resonators available so far can store photons for a few tens of microseconds only. It is thus rather difficult to process the quantum information that light carries in a quantum way. We need quantum memories.

One natural approach, when it comes to single photons, is to map them onto the state of a single atom, the interface being provided by a resonant cavity, for instance. Atom-photon information-exchange experiments have already been realized, but at the expense of using quite complex experimental techniques<sup>8</sup>. Moving to the many-photon case, mesoscopic light pulses can be 'stopped' in their tracks<sup>9</sup>: the light velocity is reduced to mere metres per second, even to zero, in an atomic medium that has an extraordinarily large index of refraction. The light field can then be mapped onto an atomic excitation and can be retrieved later (milliseconds later in real situations). But so far these experiments, although spectacular, have involved rather intense pulses of light, whose quantum properties are not apparent.

Julsgaard *et al.*<sup>2</sup> worked instead with light pulses made up of only a few photons, mapping their quantum properties onto those of an atomic ensemble. The experimental set-up is quite simple, based on glass cells holding atomic vapour of caesium at a temperature close to room temperature. For an efficient copy to be made, light and atoms must have the same quantum structure. In the special conditions used here, both are described by two continuous quantum variables, or observables —  $\hat{X}_L$  and  $\hat{P}_L$  for the laser,  $\hat{X}_A$  and  $\hat{P}_A$  for the atoms. Each pair is equivalent to the observable quantities position and momentum for the motion of a single particle. These are incompatible:



## 100 YEARS AGO

I have read with interest in your columns... a carefully compiled and instructive account of the discussions that have from time to time during the past 50 years broken out with regard to the naming of the highest measured point on the earth's surface, Peak XV of the Indian Survey. I have long maintained it to be a matter for regret that the monarch of mountains should be called after any individual, however eminent, and I am still of this opinion, which is shared by most mountaineers and mountain lovers. We should prefer that Peak XV should bear a Nepalese or a Tibetan name, even had one to be invented for it, as twenty years ago Alpine Clubmen, in accord with Russian surveyors, found or invented names for many of the great peaks of the Caucasus... Should [the Royal Geographical Society] resolve that, considering the length of time the title "Mount Everest" has been more or less in use in this country for Peak XV, the absence of any evidence that that individual peak is designated as, or included in the designation of Gaurisankar by the Nepalese, and the practical inconvenience (whether the name be authentic or not) of introducing a new Tibetan name such as Chomo- or Jamokangkar, it is expedient that the title Mount Everest should be generally accepted, I shall acquiesce.

Douglas W. Freshfield

From *Nature* 24 November 1904.

## 50 YEARS AGO

Studies on the pharmacology of extracts of *Rauwolfia serpentina*... reported that the alkaloid... had a marked hypotensive effect which was in part due to depression of central nervous system mechanisms... Our own studies have confirmed that [the *Rauwolfia* alkaloid] reserpine diminished reflex vasomotor responses, but have also demonstrated a direct effect on the peripheral vessels independent of its nervous activity... We have found that injections of reserpine into the systemic circulation of the rabbit produce an immediate fall in systemic blood pressure. This is accompanied by an immediate rise in limb perfusion pressure instead of a fall, as would have been expected were the fall of blood pressure mediated through the nervous system. Furthermore, injection of reserpine directly into the artery of the perfused hind-limb causes immediate diminution in vasomotor tone.

From *Nature* 27 November 1954.

through Heisenberg's uncertainty principle, as one is measured more precisely the information on the other is degraded. This incompatibility severely limits the fidelity of storage and retrieval of information in a simple 'classical' memory, in which  $\hat{X}_L$  and  $\hat{P}_L$  are measured and these values are then imprinted on  $\hat{X}_A$  and  $\hat{P}_A$ .

Julsgaard *et al.* have achieved a faithful copy with a more elaborate scheme, reminiscent of their recent experiment on atomic entanglement<sup>10</sup>. It proceeds in two steps. One of the light observables is first directly copied onto the atomic system through a non-resonant laser-atom interaction. The second step is then more akin to the classical memory operation: the other light observable is measured and the measurement result is fed back onto the atomic system, completing the memory operation.

To follow the process in more detail, imagine that the initial quantum properties of the input light pulse are described by  $\hat{X}_L^{\text{in}}$  and  $\hat{P}_L^{\text{in}}$ , and those of the atomic sample by  $\hat{X}_A^{\text{in}}$  and  $\hat{P}_A^{\text{in}}$ . The laser pulse crosses the cell of caesium vapour. It is not absorbed, but  $\hat{X}_A$  — which is now  $\hat{X}_A^{\text{in}} + \hat{P}_L^{\text{in}}$  — stores the  $\hat{P}_L$  information (with a bit of added noise, owing to the initial quantum uncertainty  $\hat{X}_A^{\text{in}}$ ). The laser observable  $\hat{X}_L$  is cast into  $\hat{X}_L^{\text{in}} + \hat{P}_A^{\text{in}}$ . In the next step of the process,  $\hat{X}_L$  is measured (destroying  $\hat{P}_L$ , but that is no longer important, because it is already stored in an atomic variable). The measured value is made negative and fed back onto  $\hat{P}_A$  by an electronic circuit and a magnetic field acting on the atoms. This achieves two goals at once: the initial quantum noise  $\hat{P}_A^{\text{in}}$  is cancelled and the observable is replaced by  $-\hat{X}_L^{\text{in}}$ . Finally,  $\hat{X}_L^{\text{in}}$  and  $\hat{P}_L^{\text{in}}$  are mapped onto  $-\hat{P}_A$  and  $\hat{X}_A$ , completing the storage operation.

In principle, the storage operation can be reversed and a light pulse identical to the input one can be regenerated. Julsgaard *et al.* preferred instead to measure the atomic observables with additional laser pulses. Through careful calibration of the quantum noise, and by comparing the probability distributions of the input and memory observables, they assessed the storage fidelity — it is significantly higher than the best possible performance of the 'measure and imprint' classical approach. But it is still not perfect, limited by experimental imperfections and initial quantum noise on  $\hat{X}_A$ . The latter could be combated in more elaborate versions of the experiment by preparing the atoms initially in a 'squeezed state', with considerably reduced fluctuations on  $\hat{X}_A$  (at the expense of increased ones on  $\hat{P}_A$ ). There should then be no limit to the fidelity.

This experiment suggests a basis for a quantum-information network operating with faint laser pulses. Obviously, there is much more work to do. The fidelity should be pushed up and the storage time increased above the present value of a few milliseconds.

An encouraging point is that this scheme uses rather simple elements, being based on clever ideas rather than heavy technology, and could therefore be turned into a practical device. It is also a clear demonstration that a large ensemble of atoms can be used as a quantum system — a line of research that is bound to generate many more exciting results.

Jean-Michel Raimond is in the Laboratoire Kastler Brossel, 24 Rue Lhomond, Paris 75005, France.  
e-mail: jmr@lkb.ens.fr

## Evolutionary biology

# Light on ancient photoreceptors

Thurston Lacall

Early multicellular organisms had two distinct types of photoreceptor cells, apparently with different functions. How these cells combined to form modern eyes turns out to be a complicated story.

The image-forming eyes, simple eyes (ocelli) and other photoreceptor organs of animals are structurally diverse. But their photoreceptor cells are basically of two types only — either 'ciliary' or 'rhabdomeric', depending on whether they use cilia or arrays of microvilli for light reception (Fig. 1). In a study of *Platynereis*, a marine segmented worm, published in *Science*, Arendt *et al.*<sup>1</sup> provide convincing evidence from gene-expression studies and sequence comparisons that the last common ancestor of bilaterally symmetric animals had both types. Their proposal for the functions the two performed, specifically the role of ciliary receptors in monitoring photoperiod, advances our understanding of the ancestral condition, before the origin of divergent types of advanced, image-forming eyes.

Our own eyes, like those of other vertebrates, have ciliary photoreceptors; so does the pineal 'third eye', a structure that is buried in the brain and is involved in circadian rhythmicity, and which still, in lower vertebrates, functions directly as a photoreceptor. The various ocelli and image-forming eyes of invertebrates, in contrast, are rhabdomeric. This, for a while, provided a useful general rule that, along with embryological differences, distinguished between the two main groups of animals: protostomes (diverse worms, molluscs and arthropods) use rhabdomeric photoreceptors; deuterostomes (vertebrates and their kin) have ciliary ones.

The person most closely associated with the idea of a dichotomy is the late Richard Eakin of the University of California, Berkeley, who carried out an extensive study of comparative eye structure using the then relatively new technique of electron microscopy<sup>2</sup>. Exceptions to the rule do occur, and in some marine flatworm larvae even

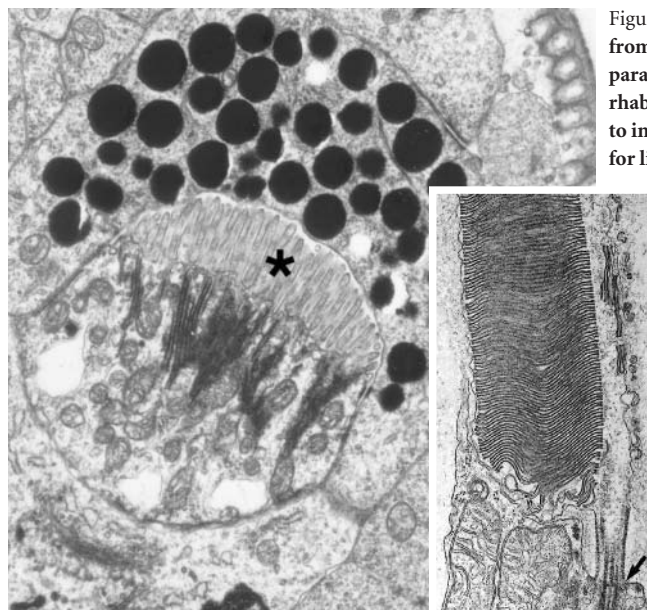
1. Bennett, C. H. & Di Vincenzo, D. P. *Nature* **404**, 247–255 (2000).
2. Julsgaard, B., Sherson, J., Cirac, J. I., Fiurásek, J. & Polzik, E. S. *Nature* **432**, 482–486 (2004).
3. Gisin, H., Ribordy, G., Tittel, W. & Zbinden, H. *Rev. Mod. Phys.* **74**, 145–195 (2002).
4. Bouwmeester, D. *et al.* *Nature* **390**, 575–579 (1997).
5. Zeilinger, A. *Rev. Mod. Phys.* **71**, S288–S297 (1999).
6. Grosshans, F. *et al.* *Nature* **421**, 238–241 (2003).
7. Furusawa, A. *et al.* *Science* **282**, 706–709 (2003).
8. Raimond, J. M., Brune, M. & Haroche, S. *Rev. Mod. Phys.* **73**, 565–582 (2001).
9. Lukin, M. D. *Rev. Mod. Phys.* **75**, 457–472 (2003).
10. Julsgaard, B., Kozhekin, A. & Polzik, E. S. *Nature* **413**, 400–403 (2001).

single ocelli can have both types of receptor cell<sup>3</sup>. The small size and sporadic occurrence of such structures has discouraged any systematic study of their function, however, so they have remained little more than anomalies in an otherwise broadly accepted general pattern.

The advantage of molecular techniques, as applied to the problem by Arendt *et al.*<sup>1</sup>, is twofold: first, their ability to reveal gene-expression patterns in individual cells; second, the inferences one can make regarding function based on the known function of homologous genes (orthologues) in other animals. The results show that there are two forms of the gene for the photopigment opsin in *Platynereis*, one ciliary, previously unknown from protostomes, and one rhabdomeric. The former is expressed in two small clusters of apical cells with internalized cilia, located in the developing brain. The cells also express an orthologue of the *rx* gene, an upstream controller of ciliary photoreceptor differentiation in vertebrates, and either they or adjacent cells show rhythmic expression of a *bmal/cycle* gene, a key component of the circadian clock. An unanswered question is the relation between the cells that express these genes and larval apical tuft cells, which are internalized intact during development in some marine worms<sup>4</sup>. An assortment of other, possibly related structures — apical ciliated pits and ampullary organs — also occur in molluscan larvae<sup>5</sup>.

Assuming that *Platynereis* does indeed preserve something of the ancestral condition (that is, of the common ancestor of protostomes and deuterostomes), the results are best explained by an early origin of two separate types of photoreceptor. Rhabdomeric ones would have been used for monitoring light direction, and ciliary ones for photoperiod. As image-forming eyes evolved,





**Figure 1 Rhabdomeric ocellus from a *Platynereis* larva. Here, a parallel array of microvilli (the rhabdome, asterisk) is employed to increase the surface available for light reception. In contrast (inset), the vertebrate eye contains ciliary photoreceptors, in which membranous lamellae are stacked within the body of a modified cilium, the base of which is arrowed. Arendt *et al.*<sup>1</sup> show that *Platynereis* contains two forms of the gene for the photopigment opsin: a rhabdomeric one, as expected, but also a ciliary one.**

which would have been a later event probably coincident with the Cambrian radiation<sup>6</sup>, some 540 million years ago, protostomes and deuterostomes followed different paths, coopting rhabdomeres and cilia respectively as visual receptors. The presence of a remnant of the ciliary system in *Platynereis* suggests that such cells may occur more widely among invertebrates than is currently recognized, which should provide an impetus for further investigations.

Do vertebrates have any remnant of the rhabdomeric system? Arendt *et al.* suggest that they do, proposing retinal ganglion cells, the neurons whose fibres form the optic nerve, as the best candidates. They also suggest that homology explains similarities in the pattern of optical pathways in protostome and deuterostome brains, but that is much more speculative.

Speculations aside, the new work provides a solid starting point for further study of the evolution of photoreceptor organs during the diversification of bilateral

multicelled animals. In evolutionary terms, it is a long way from a simple ocellus, involving no more than a few cells, to the complexity of an optimally constructed image-forming eye. Evolution seems to have accomplished this transition piecemeal, by myriad small steps, each an adaptive improvement over what went before<sup>7</sup>. A detailed accounting of the steps is as yet beyond us, but clarifying the nature of the ancestral condition is a useful beginning. ■

Thurston Lacalli is at the Department of Biology, University of Victoria, Victoria, British Columbia V8W 3N5, Canada.

e-mail: lacalli@uvic.ca

1. Arendt, D., Tessmar-Raible, K., Snyman, H., Dorreesteijn, A. W. & Wittbrodt, J. *Science* **306**, 869–871 (2004).
2. Eakin, R. M. in *Visual Cells in Evolution* (ed. Westfall, J. A.) 91–105 (Raven, New York, 1982).
3. Eakin, R. M. & Brandenberger, J. L. *Science* **211**, 1189–1190 (1981).
4. Lacalli, T. C. *Proc. R. Soc. Lond. B* **212**, 381–402 (1981).
5. Page, L. R. & Parries, S. C. *J. Comp. Neurol.* **418**, 383–401 (2000).
6. Parker, A. R. *Proc. R. Soc. Lond. B* **265**, 967–972 (1998).
7. Lenski, R. E. *et al. Nature* **423**, 139–144 (2003).

## Nonlinear physics

# Fresh breather

David K. Campbell

The direct observation of highly localized, stable, nonlinear excitations — known as discrete breathers — at the atomic level underscores their importance in physical phenomena at all scales.

Stanislaw Ulam, the celebrated Polish mathematician and godfather of the field now known as nonlinear science, famously remarked that using the term ‘nonlinear science’ was like ‘calling the bulk of zoology the study of non-elephants’. He meant that linear processes are the exception rather than the rule; that most phenomena

are inherently nonlinear; and that the effects of nonlinearity are apparent everywhere in nature, from the synchronized flashing of fireflies through clear-air turbulence to tornadoes and tsunamis.

Perhaps Ulam should have carried the metaphor of a nonlinear ‘zoo’ a bit further, for the remarkable taxonomy of ‘non-elephants’

observed in recent physics experiments has been truly breathtaking. One of the newest species to be added to the menagerie is the ‘discrete breather’, or intrinsic localized mode. Briefly characterized, discrete breathers are spatially localized, time-periodic, stable excitations that exist and propagate in spatially extended, perfectly periodic, discrete systems. On page 486 of this issue, Sato and Sievers<sup>1</sup> report the sighting of a particularly elusive form of discrete breather that exists at the atomic scale in a magnetic solid. Coupled with other recent observations in systems ranging from Josephson-junction arrays<sup>2,3</sup>, through micromechanical systems<sup>4</sup>, to photonic crystals<sup>5</sup> and optical-switching waveguide arrays<sup>6</sup>, this new observation underscores the ubiquity of these nonlinear excitations and the importance of understanding their role in determining the properties of these widely different physical systems<sup>7</sup>.

To understand Sato and Sievers’ elegant experiment and its implications, it is best to start with the core principles of solid-state physics at the linear level, and then build up (to use a deliberate oxymoron) some counter-intuitive ‘nonlinear’ intuition. At the atomic level, crystalline solids are made up of discrete arrays — lattices — of atoms (or molecules) that typically form a regular, periodic structure, like the sleepers (ties) supporting railway tracks. As a result, linear excitations — be they electrons or phonons (the particles that carry sound) — moving through a solid will experience a periodic energy potential. This implies, by general mathematical theorems devised by Bloch and Floquet, the existence of ‘forbidden’ and ‘allowed’ bands of frequency and velocity for their motion. Linear excitations can propagate through the solid only in the allowed bands, which cover a finite range of frequencies and thus have a highest and a lowest allowed frequency. Importantly, the discreteness of the lattice determines this highest frequency; to continue the analogy, the farther apart the railway sleepers, the smaller the highest frequency allowed.

What are the consequences of nonlinearity? Indeed, what does ‘nonlinear’ mean in this context? To answer these questions, let us recall one of the hallmarks of the dawn of modern science: Galileo’s observation of the oscillations of the censers in the cathedral in Pisa. This is now replicated in all elementary physics courses as the ‘plane pendulum experiment’. It is a classic example in at least two senses. First, it is a problem that all novice physicists solve. Second, it is a typical example of how we mislead ourselves and our students about the prevalence and importance of nonlinearity. Without going into the mathematics, just remember that the result we are supposed to discover from this experiment is that the frequency of oscillation of the pendulum does not depend on its amplitude (that is, how far the censer

swings) but only on the length of the pendulum and the acceleration due to gravity. In fact, this result is an approximation, true only when the pendulum undergoes small (linear) oscillations. If the pendulum makes large excursions from the vertical, its frequency actually decreases with its amplitude: larger deviations have longer periods (and therefore slower frequencies). So in this 'nonlinear regime', the frequency of the pendulum does depend on its amplitude.

We can now understand how discreteness and nonlinearity lead to the existence of stable discrete breathers. The discreteness of the system means that there is a finite range of frequencies in which linear excitations can exist. If a large-amplitude (and hence nonlinear) excitation is created — a putative discrete breather — its frequency can be shifted out of the allowed band of linear excitations; for nonlinear systems such as the plane pendulum (with so-called soft nonlinearities), this frequency lies below the lowest allowed linear frequency. For highly discrete systems, the allowed linear band can be very narrow, as the highest allowed linear frequency decreases with increasing discreteness. This means that harmonics of the discrete-breather frequency can all lie above the linear band, so that the discrete breather cannot couple to linear excitations and is therefore stable against decaying into them. Although there are additional subtleties<sup>7</sup>, these simple arguments capture the essence of the phenomenon.

In their experimental study, Sato and Sievers<sup>1</sup> use a quasi-one-dimensional anti-ferromagnetic solid. One can consider this as a chain-like material in which the individual units are tiny spins, which point alternately up and down. The linear excitations in this discrete, spatially extended system are known as 'spin waves' and, as anticipated, exist only in a finite band (or spin-wave manifold). The discrete breathers can be envisaged as excitations that are highly localized (involving only a few spins) but highly excited (rapidly precessing, hence in the nonlinear regime), with frequencies lying below (and harmonics lying above) the spin-wave manifold. The small number of spins in the discrete breather makes direct observation in, say, an absorption experiment essentially impossible.

To circumvent this difficulty, Sato and Sievers<sup>1</sup> use a novel nonlinear spectroscopic technique that enhances the signal produced by the discrete breathers and suppresses the linear spin-wave signal (through a third-order nonlinearity in the magnetic susceptibility of the material). This enables them to 'lock' a small number of discrete breathers to an external driving frequency — and as the frequency of the discrete breather depends on its amplitude, a given external frequency selects a particular amplitude for the excitation. Then, by tuning a third mixing frequency, the individual discrete breathers

can be unlocked essentially one at a time. This provides striking confirmation of the existence, as well as the nature, of the individual nonlinear excitations.

Given the ubiquity of such breathers in discrete nonlinear physical systems (which exist on essentially all length scales), these nonlinear excitations are likely to be important in many physical phenomena, including melting, fracture, and the buckling and folding of biopolymers. They may also prove useful in technologies ranging from 'smart' materials with tunable collective responses to light-induced, all-optical switches and networks. With the acquisition of this new animal, the nonlinear 'zoo' has become

an altogether more interesting place.

David K. Campbell is at the College of Engineering, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA.

e-mail: [dkcampbe@bu.edu](mailto:dkcampbe@bu.edu)

1. Sato, M. & Sievers, A. J. *Nature* **432**, 486–488 (2004).
2. Trias, E., Mazo, J. J. & Orlando, T. P. *Phys. Rev. Lett.* **84**, 741–744 (2000).
3. Binder, P., Abaimov, D., Ustinov, A. V., Flach, S. & Zolotarev, Y. *Phys. Rev. Lett.* **84**, 745–748 (2000).
4. Sato, M., Hubbard, B. E., Sievers, A. J., Ilic, B. & Craighead, H. G. *Europhys. Lett.* **66**, 318–323 (2004).
5. Fleischer, J. W., Segev, M., Efremidis, N. K. & Christodoulides, D. N. *Nature* **422**, 147–150 (2003).
6. Eisenberg, H. S., Silberberg, Y., Morandotti, R., Boyd, A. R. & Aitchison, J. S. *Phys. Rev. Lett.* **81**, 3383–3386 (1998).
7. Campbell, D. K., Flach, S. & Kivshar, Yu. S. *Phys. Today* 43–48 (January 2004).

## Molecular biology

# Termination by torpedo

David Tollervey

The information encoded in our genes must be copied into messenger RNAs, which will programme the protein-synthesis machinery. New results support an intriguing mechanism for ending the copying process.

Genetic information is encoded in the sequence of the DNA double helix. To access this information, the DNA sequence must be copied, or 'transcribed', by enzymes known as RNA polymerases. The resulting messenger RNA (mRNA) molecules carry the genetic information to the protein-synthesizing machinery, where it is used to define the amino-acid sequence, and therefore the structure and function, of proteins. A lot is known about how transcription is started, but the mechanisms that stop the transcription of mRNA by RNA polymerase II are less well understood. Transcription termination is important, not least because stopping too late will disrupt the regulation of other genes on the same chromosome; stopping prematurely, meanwhile, would produce truncated and therefore defective mRNAs. So it initially seemed likely that there would be strictly defined 'stop' sequences. But termination on most genes has been found to occur at various positions rather than at a single site. Moreover, a discrete signal in the DNA that might define the termination site could not be identified.

Three papers in this issue<sup>1–3</sup> address this problem and provide evidence in favour of a striking alternative mechanism: that the transcribing RNA polymerase is 'torpedoed'. The mRNA is cut while it is still being synthesized, with the liberated region forming the mRNA. The remaining RNA trails out of the transcribing RNA polymerase, and is attacked by an exonuclease — an enzyme that can degrade the RNA from its free end. This enzyme chases after the polymerase, chewing up the RNA strand as it goes. When it catches up with the polymerase, transcription is terminated.

To understand the details of the new findings, a bit more background is necessary. A defining feature of mRNAs is a tail composed of a long string of adenosine nucleotides — the poly(A) tail. This is not encoded by the gene but is added following cleavage of the nascent RNA transcript. Molecular factors that recognize the cleavage site, and cut the RNA, bind to a regulatory region of the transcribing RNA polymerase II called the carboxy-terminal domain (CTD). This interaction is important for recruiting the factors to the nascent transcript. In turn, these factors must be off-loaded from the polymerase onto the RNA at their site of action for transcription to be terminated<sup>4–7</sup>. In fact, recognition of the cleavage site has been believed to be the key step in termination — but the new data<sup>1–3</sup> show that this isn't the whole story.

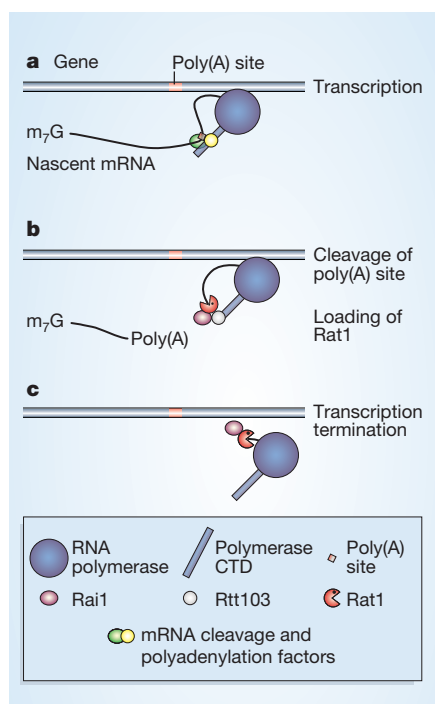
Working with yeast, Kim *et al.*<sup>1</sup> identified Rtt103 as a protein that can bind to a peptide — a short stretch of amino acids — from the CTD. They then showed that Rtt103 is associated with the exonuclease Rat1 and its cofactor Rai1 (Fig. 1). They hypothesize that Rat1 is recruited to transcripts by Rtt103 bound to the CTD of the transcribing polymerase, and independently by Rai1. They also found that inactivating Rat1 in mutant yeast strains strongly stabilized the RNA fragment downstream from the cleavage/polyadenylation site, indicating that Rat1 is responsible for degrading this region. Moreover, RNA polymerase II could be chemically crosslinked to sequences in the encoding gene that are farther downstream from the cleavage site in the mutants than in the wild type, suggesting that, without functional



Rat1, the polymerase fails to stop when it should.

The authors conclude that when the mRNA is cut at the polyadenylation site, Rat1 is recruited to the free end. Rat1 is highly processive, so that once bound it will rapidly degrade the RNA strand until it catches the transcribing polymerase. This, the authors propose, triggers termination. Kim *et al.* also extended their studies to an entire chromosome, and show that normal transcription termination on many mRNA-encoding yeast genes requires the activity of Rat1.

Broadly similar conclusions were arrived at by Teixeira *et al.*<sup>2</sup> and West *et al.*<sup>3</sup>, but from a very different starting point. Previous



**Figure 1 Transcription termination in yeast.** a, Transcription of a gene by RNA polymerase produces a nascent messenger RNA (mRNA). mRNAs eventually have two distinctive features: a 7-methyl-guanosine cap (m<sub>7</sub>G) at one end and a polyadenosine (poly(A)) tail at the other. The carboxy-terminal domain (CTD) of the polymerase comprises repeats of a seven-amino-acid motif and undergoes reversible modification by phosphorylation (reviewed in ref. 11; not shown). In particular, phosphorylation of a serine amino acid at position 2 in each repeat increases as the polymerase moves along the mRNA. b, Kim *et al.*<sup>1</sup> have found that the Rtt103 protein binds a CTD peptide carrying the serine 2 phosphorylation, and also associates with the exonuclease Rat1 (Xrn2 in humans) and its cofactor Rai1. Thus, Rat1 may be recruited to the nascent mRNA via Rtt103 bound to the polymerase CTD, and by an independent system that involves Rai1. Following cleavage of the poly(A) site, Rat1 binds and eats away at the free end of the RNA, halting transcription when it catches the polymerase (c).

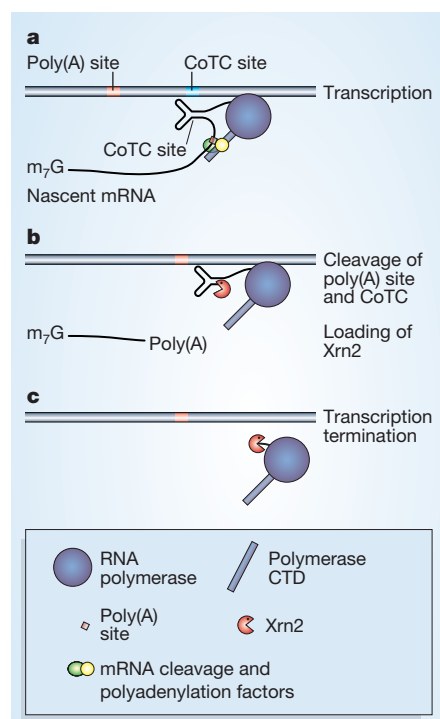
analyses<sup>8</sup> of the human  $\beta$ -globin gene identified two mRNA cleavage sites: the normal polyadenylation site and another site, the co-transcriptional cleavage (CoTC) site (Fig. 2). As its name suggests, the CoTC is cleaved in nascent transcripts and is required for efficient transcription termination. Teixeira *et al.* show that, surprisingly, the RNA sequence at the CoTC forms an RNA enzyme (a ribozyme) with a structure that has intrinsic self-cleavage activity in the absence of proteins. Mutations in the CoTC that inhibit self-cleavage also inhibit transcription termination.

West *et al.* add to these observations by showing that depletion of Xrn2 (the human counterpart of yeast Rat1) using RNA interference also impairs transcription termination. Notably, termination depended on the presence of the polyadenylation site, in addition to the CoTC. This suggests that recognition of the polyadenylation site by the processing factors associated with the polymerase CTD makes it competent for subsequent termination by Rat1/Xrn2.

Polyadenylation sites are present in all genes that encode mRNAs, whereas the CoTC element has been identified only in primate  $\beta$ -globin genes. This suggests that cleavage at the polyadenylation sites is the normal means of entry for Rat1/Xrn2. The globin genes are very highly transcribed in some cell types, and so the CoTC may have evolved as a means of ensuring their efficient termination. However, although the CoTC RNA has features in common with other self-cleaving RNA enzymes, it could not have been recognized as a potential ribozyme from its sequence alone<sup>2</sup>. So other, as-yet-uncharacterized ribozymes might lie downstream of many genes.

The emerging picture of transcription termination is of a race between the Rat1/Xrn2 nuclease and the elongating polymerase. Cleavage of the nascent mRNA at the CoTC lets the nuclease 'jump ahead' by providing a downstream entry site. Similar effects would presumably result from an alternative site, one that slows the polymerase, allowing Rat1 to catch it more rapidly. Polymerase pause sites have indeed been implicated in transcription termination in other systems.

Termination may not, however, require highly specific interactions between the Rat1/Xrn2 nuclease and the polymerase, as relocating another nuclease, the cytoplasmic protein Xrn1, to the nucleus can suppress the lethality of a Rat1 mutation in yeast<sup>9</sup>. And this finding provides a potential connection to another model: in the bacterium *Escherichia coli*, the transcription termination factor Rho tracks along nascent transcripts and physically pulls the RNA away from the polymerase when this is slowed by pause sites (reviewed in ref. 10). It is conceivable that a highly processive exonuclease,



**Figure 2 Transcription termination at the  $\beta$ -globin gene in human cells.** a, The newly synthesized human  $\beta$ -globin mRNA will be cut at two positions: at the normal site of poly(A) addition and within the co-transcriptional cleavage (CoTC) site, which Teixeira *et al.*<sup>2</sup> show to be self-cleaving. b, West *et al.*<sup>3</sup> propose that this allows the entry of Xrn2 (the human Rat1 counterpart), which again eats away at the mRNA until it catches up with the RNA polymerase (c). Replacing the CoTC region with a different self-cleaving RNA, a 'hammerhead ribozyme', inhibits transcription termination, possibly reflecting specific recruitment of Xrn2 to the CoTC<sup>3</sup>. However, the hammerhead cleavage leaves a product with a 5'-hydroxyl group, which is a poorer substrate for Rat1 than the 5'-phosphate produced by cleavage of the CoTC, so it might be a poorer substrate for Xrn2 as well.

such as Rat1, would also pull on the nascent RNA transcript if physically impeded by the polymerase. In any event, the new observations<sup>1-3</sup> will open the way to more mechanistic analyses of transcription termination than have hitherto been possible.

David Tollervey is at the Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3JR, UK.  
e-mail: d.tollervey@ed.ac.uk

- Kim, M. *et al.* *Nature* **432**, 517–522 (2004).
- Teixeira, A. *et al.* *Nature* **432**, 526–530 (2004).
- West, S., Gromak, N. & Proudfoot, N. J. *Nature* **432**, 522–525 (2004).
- Aranda, A. & Proudfoot, N. *Mol. Cell* **7**, 1003–1011 (2001).
- Barilla, D., Lee, B. A. & Proudfoot, N. J. *Proc. Natl Acad. Sci. USA* **98**, 445–450 (2001).
- Dichtl, B. *et al.* *EMBO J.* **21**, 4125–4135 (2002).
- Meinhart, A. & Cramer, P. *Nature* **430**, 223–226 (2004).
- Dye, M. J. & Proudfoot, N. J. *Cell* **105**, 669–681 (2001).
- Johnson, A. W. *Mol. Cell Biol.* **17**, 6122–6130 (1997).
- Kaplan, D. L. & O'Donnell, M. *Curr. Biol.* **13**, R714–R716 (2003).
- Burawski, S. *Nature Struct. Biol.* **10**, 679–680 (2003).

## Immunology

### Added piece for 'bubble boy' puzzle

*J. Clin. Invest.* **114**, 1512–1517 (2004)

Severe combined immunodeficiency (SCID) is a life-threatening disease characterized by a crippled immune system. It is often referred to as 'bubble boy disease', after David Vetter, a SCID sufferer who spent 12 years living inside a protective plastic bubble. The condition can be caused by numerous inherited mutations; geneticists have now identified another one.

Geneviève de Saint Basile *et al.* report that patients and affected fetuses from three families have mutations in one of the genes that encode subunits of the CD3 receptor. This protein complex is found on T cells, one arm of the immune response to intruders. In particular, mutations in either the *CD3E* or the *CD3D* gene cause a complete lack of T cells, showing that the products of these genes are required for a crucial step in T-cell development.

Researchers had previously identified other immune defects that cause the disease, including a lack of the  $\gamma$ c chain of the interleukin-2 receptor, a protein essential for T cells and another immune-cell type, NK cells, to develop. The patients with CD3 mutations had normal NK cells, however, underlining the range of causes of this disease. Screening for these mutations might, the authors suggest, open up new possibilities for gene therapy. **Michael Hopkin**

## Granular media

### Why nuts come in pairs

*Phys. Rev. Lett.* **93**, 208002 (2004)

Everyone knows that Brazil nuts in a shaken box of muesli rise to the top. But large grains dispersed among smaller grains can also sink to the bottom when shaken: the 'reverse Brazil-nut effect'. Which type of behaviour is observed depends on factors such as the relative sizes and densities of the particles.

Duncan A. Sanders *et al.* have found a third type of behaviour in a vibrated mixture of differently sized grains. If the large grains are neutrally buoyant — that is, have about the same density as the small grains — they neither rise nor sink, but aggregate into a cluster. In other words, the mixture separates out into regions of small and large grains. The researchers' computer simulations show that there is effectively a force of attraction between the large grains, extending to about five large-grain diameters.

The attraction seems to result from the way the packing of smaller 'host' grains dilates during a vibration cycle, leading to more particle collisions on the outward-facing sides of the large grains than in the

## Neurobiology

### Cells take sides on vision

*Cell* **119**, 567–578 (2004)

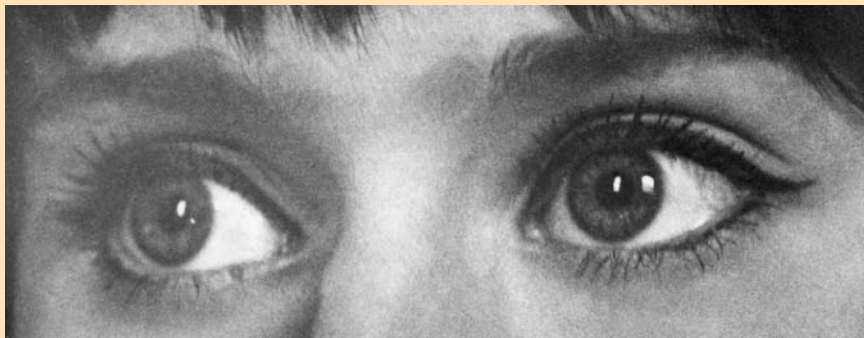
Thanks to a carefully integrated set of neurons, we can combine what our two eyes see to form one picture. This binocular vision hinges on the pathfinding decisions made during development by the axons of ganglion cells in the retina that go on to form the optic nerve. But how do these cells choose between visual centres on the left or right side of the brain?

Winnie Pak *et al.* have tackled this question with genetics; they find that the gene-transcription factor *Islet-2* is crucial in determining the laterality of retinal ganglion cells in mice. More

specifically, they show that this protein directs the cells from one eye to project towards a region on the opposite side of the brain. In mice lacking *Islet-2*, an abnormally large number of retinal ganglion cells prefer to keep to their side of origin. The findings indicate that two genetically unique domains make up the retina — and that these domains decide the laterality of specific cells involved in sight.

Pak *et al.* emphasize that a genetic hierarchy is responsible for proper development of binocular vision. They suggest that, under normal circumstances, *Islet-2* works by repressing *Zic2*, another transcription factor, which activates a one-sided axon pathfinding programme.

**Roxanne Khamsi**



region between them. The large grains are thus gradually 'ratcheted' together by the vibrations, an effect that could occur during the industrial handling of powders. **Philip Ball**

## Organic chemistry

### Look, no metal

*Angew. Chem. Int. Edn* doi:10.1002/anie.200461816 (2004)

There are plenty of circumstances in which synthetic chemists need to modify an organic molecule at a specific position without affecting another part of that molecule. They can use a catalyst, but for some reactions catalysts can be expensive, contain environmentally harmful transition metals or require high temperatures.

One common chemical reaction requires conversion of a carbon–carbon double bond to a single bond in the presence of an adjacent carbon–oxygen double bond. Jung Woon Yang *et al.* report a simple way of carrying out this conversion without the metal catalyst or harsh reaction conditions that are usually required. Technically, this is an olefin reduction: in Yang and colleagues' scheme it involves the use of an amine catalyst and a nicotinamide adenine dinucleotide-like hydrogen donor, an agent that certain enzymes employ in the natural world.

The authors say that the reaction is highly selective and more environmentally

friendly than other methods. Furthermore, in another paper in press at the same journal, they describe how they have applied their procedure to selectively produce one of two possible stereoisomers of the same molecule — often a highly desirable end. **Joshua Finkelstein**

## Applied physics

### Transistor makes light work

*Appl. Phys. Lett.* **85**, 4768–4770 (2004)

Transistors are one of the fundamental components in any electronic circuit. They are essentially three-terminal switches: a current at one port controls the current between the others. G. Walter *et al.* have now created a transistor where the output is laser light as well as electrical current. This could speed up the conversion of electrical signals from a conventional circuit into light that can travel along a fibre-optic cable.

Their device sandwiches together indium gallium phosphide, gallium arsenide and indium gallium arsenide. This last material acts as a 'quantum well' that captures electrons, leading to the generation of coherent laser light.

Operating at about  $-73^\circ\text{C}$ , the device begins to emit infrared laser light when about 8 mA of current flows through it — although the authors note that the efficiency of the unit drops sharply when it switches from non-coherent light emission to lasing. They are now working to optimize this output. **Mark Peplow**



# Groundwater maintains dune landscape

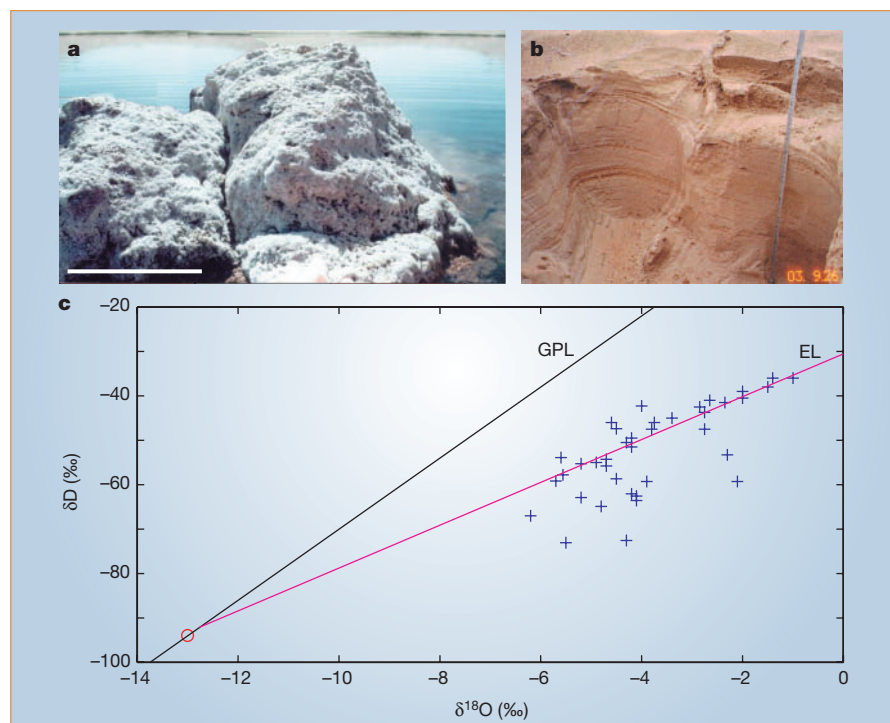
A remote water source helps giant sand dunes to stand their ground in a windy desert.

The Badain Jaran desert in western Inner Mongolia in China has a unique landscape that contains 72 lakes, with a total water surface area of 23 km<sup>2</sup>, and the world's highest stationary sand dunes, which are up to 500 m tall — despite the prevailing dry and windy weather conditions<sup>1,2</sup>. Here we present evidence of a major groundwater system that underpins the factors leading to this landscape. Our finding could transform plans for the region's water resources.

During two visits to the desert (for map of study site, see supplementary information) in 2003, we discovered that the large sand dunes were relatively moist underneath a dry surface layer of about 20 cm. At the top of the dunes, the moisture content measured over a depth of 2 m from the surface layer occupied 2–20% of the space between sand grains, known as the pore volume. Seepage of pore water was observed in a one-metre-deep well that was dug on the flank of a dune, even though the well was 17 m above the water level in a nearby lake. This water is likely to be acting as a cohesion agent, providing the dunes with resistance against wind erosion and transportation.

A large disparity exists in the desert between the local rainfall (about 40 mm a year) and net evaporation rates (about 200 mm and 4,000 mm a year for the dune and lake areas, respectively)<sup>2,3</sup>. We found evidence that the dune and lake water originates from snowmelt on Qilian Mountain, which lies 500 km southwest of the desert. The snowmelt percolates into the mountain's Deep Big Faults system<sup>1</sup>, which in turn feeds deep carbonate layers that extend from the mountain to the desert (see supplementary information); leakage through fractures of the deep carbonate layers then delivers water to the lakes and dunes in the desert.

Calc-sinters (crystalline deposits of calcium carbonate) and rhizoconcretions (fossilized plant roots coated in calcium carbonate) were found in the lakes and dunes (Fig. 1a,b). Using <sup>14</sup>C-dating, we found that an undisturbed rhizoconcretion buried inside a dune was 4,753 ± 60 years old, indicating that the dune had been stationary over this time. Calc-sinters and rhizoconcretions are formed from carbonate-rich groundwater<sup>4</sup>. Strontium-isotope analysis indicates that groundwater passing through the carbonate layers provides the calcium in the samples. The <sup>87</sup>Sr/<sup>86</sup>Sr ratio of the samples ranges from 0.710 to 0.713, which is close to the range of 0.705–0.709 found for carbonates resulting from sea sedimentation 250–600 Myr ago<sup>5</sup>. The difference might be due to contact of



**Figure 1** Evidence for a groundwater system under the Badain Jaran desert. **a**, Calc-sinter found in Yihejigede Lake (longitude, 102° 09.498'; latitude, 39° 46.099'; altitude, 1,143 m). Spring flow was measured in the middle of the calc-sinter as 1 litre per second. Scale bar, 1 m. **b**, Buried rhizoconcretion (longitude, 102° 09.741'; latitude, 39° 45.642'; altitude, 1,160 m). The depth of the excavated sand layer is about 30 cm. **c**, Isotope  $\delta D$ – $\delta^{18}O$  composition of snow from Qilian Mountain (circle) and desert groundwater (crosses); GPL, global precipitation line; EL, evaporation line. The snow data are taken from previously published work<sup>6</sup>.

groundwater with overlying <sup>87</sup>Sr-enriched rocks during upwelling from the deep carbonate layers through fractures.

We analysed groundwater samples collected from deep wells and springs in the desert for their  $\delta D$  and  $\delta^{18}O$  composition. The results follow the evaporation line plotted in Fig. 1c, which intersects the global precipitation line near the data point of  $\delta D$ – $\delta^{18}O$  measured for snow from the Faults area. This isotopic connection provides evidence that the groundwater is sourced from the snow, indicating that the snowmelt from Qilian Mountain recharges the groundwater system underneath the desert. Measurements of <sup>3</sup>H and three different forms of chlorofluorocarbons (CFC-11, 12 and 13) give the age of the spring water from the desert as between 20 and 30 years old. The relatively young water age is consistent with the fast passage of water through the Faults and deep carbonate layers.

The desert groundwater system represents a major water resource for the region. In the regional water balance, groundwater recharge equals evaporation minus rainfall (see supplementary information). We used this relation to estimate the groundwater yield of the Badain Jaran desert, finding it to be  $5 \times 10^8$  m<sup>3</sup> yr<sup>-1</sup>. A large water-diversion

project, with an annual capacity of  $2.5 \times 10^8$  m<sup>3</sup>, is being planned for areas north of Qilian Mountain at a cost of more than US\$500 million. The groundwater resource identified here could provide a practical alternative to the diversion project. However, its utilization would have to be undertaken with care because any resulting dune mobilization could severely affect the regional eco-environment.

**Jian Sheng Chen\***, **Ling Li†‡**, **Ji Yang Wang\*§**, **D. A. Barry||**, **Xue Fen Sheng¶**, **Wei Zu Gu\***, **Xia Zhao\***, **Liang Chen\***

\*Institute of Isotope Research, and †Centre for Eco-Environmental Modelling, Hohai University, Nanjing, 210098, China

‡School of Engineering, University of Queensland, St Lucia, Queensland 4072, Australia  
e-mail: l.li@uq.edu.au

§Institute of Geology and Geophysics, Chinese Science Academy, Beijing 100018, China

||School of Engineering and Electronics, University of Edinburgh, Edinburgh EH9 3JL, UK

¶Institute of Surficial Geochemistry, Department of Earth Sciences, Nanjing University, Nanjing 210093, China

1. Yang, X.-P. *Quat. Sci.* **22**, 97–104 (2002).

2. Yang, X.-P. *Sci. Rep.* **45**, 428–434 (2000).

3. Huang, C.-Y. *Clim. Envir. Res.* **18**, 84–90 (2003).

4. Liu, Z. *et al. Sci. China D* **43**, 569–576 (2000).

5. Huang, S.-J. *et al. Adv. Earth Sci.* **1**, 194–200 (2001).  
6. Ren, J.-W. *J. Glaciol. Geocryol.* **21**, 151–154 (1999).  
Supplementary information accompanies this communication on Nature's website.  
Competing financial interests: declared none.

# Palaeoclimate

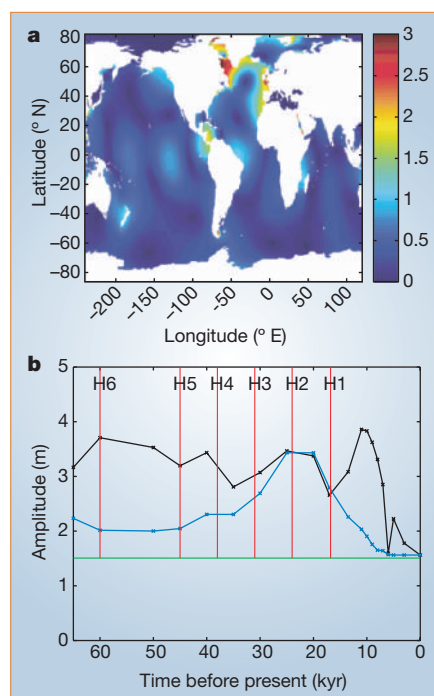
## Ocean tides and Heinrich events

Climate varied enormously over the most recent ice age<sup>1</sup> — for example, large pulses of ice-rafted debris<sup>2</sup>, originating mainly from the Labrador Sea<sup>3</sup>, were deposited into the North Atlantic at roughly 7,000-year intervals, with global climatic implications<sup>3</sup>. Here we show that ocean tides within the Labrador Sea were exceptionally large over the period spanning these huge, abrupt ice movements, which are known as Heinrich events. We propose that tides played a catalytic role in liberating iceberg armadas during that time.

We investigated whether tides could have been linked to Heinrich events for two reasons. First, tidal controls on continental ice streams and floating ice shelves, both proposed<sup>4,5</sup> as sources of Heinrich icebergs, are well documented<sup>6,7</sup> in present-day Antarctica. Second, ice-age tides should differ from those of today. The growth of continental ice sheets was accompanied by lower globally averaged sea levels (up to 130 m)<sup>8</sup>, with an implied decrease in tidal phase speeds and in ocean-basin size, both of which affect the strong resonance<sup>9</sup> of North Atlantic semidiurnal tides. Lower sea levels also affected tides by reducing the area of shallow-water regions, where much of the dissipation of present-day tides takes place.

We predicted ice-age tides in a global numerical model<sup>10</sup> that captures 92% of the present-day open-ocean tidal-height variance. The ice-age simulations required, as input, the space–time history of the ice-sheet distribution and of the complex global geometry of sea-level variations. The latter fields were generated using a formulation<sup>11</sup> for predicting gravitationally self-consistent sea-level changes on viscoelastic Earth models. (For details of tide and sea-level models, see supplementary information.)

Figure 1a shows the modelled amplitudes of  $M_2$ , the largest tidal constituent, 45,000 years ago (45 kyr). The Labrador Sea amplitude (about 3.2 m) is much larger than in other deep areas of either the 45-kyr or present-day ocean. The black line in Fig. 1b shows the predicted  $M_2$  amplitude at 61.5° N, 64° W, the approximate discharge point of the Hudson Strait ice stream<sup>5</sup>, over the past 65 kyr. All simulations spanning the Heinrich events (H1–H6), and extending to about 7 kyr, predict amplitudes of 2.7–3.9 m; such amplitudes are roughly twice the present-day value (about 1.5 m; ref. 12). Experiments with many



**Figure 1** Ice-age tidal amplitudes. **a**, Amplitude (m) of the principal lunar semidiurnal tide  $M_2$  at 45,000 years ago (45 kyr) in a hydrodynamical model<sup>10</sup> coupled to a gravitationally self-consistent (hence geographically variable) prediction<sup>11</sup> of sea-level change. **b**, Black line:  $M_2$  amplitude (m) from the same models, at the estimated discharge point of the Hudson Strait ice stream, over the past 65 kyr. (See supplementary information for discussion of uncertainties.) Blue line:  $M_2$  amplitudes at the same Labrador Sea location in simulations using globally averaged sea-level change applied in a spatially uniform manner. In this calculation, ocean-basin geometry is altered by the uniform sea-level change, but not by any growth or collapse of marine-based ice. Green line: present-day  $M_2$  amplitude at the same location in a very accurate satellite-constrained tide model<sup>12</sup>. Timings of Heinrich events H1–H6 are from ref. 3.

tidal frequencies yield maximum peak-to-peak ranges over the spring–neap cycle that are 3.6 times larger than the  $M_2$  amplitudes, or 10–14 m. This greatly exceeds the maximum ranges in present-day Antarctica, where tides are thought to weaken floating ice shelves by forming crevasses at their hinge lines.

What feature of the ice-age ocean amplifies Labrador Sea tides? The blue line in Fig. 1b shows Labrador Sea  $M_2$  amplitudes in simulations for which the globally averaged sea-level change is applied in a spatially uniform manner. Furthermore, the ocean-basin geometry is assumed to be influenced by the sea-level changes but not by changes in the ice geometry. In this case, large Labrador Sea tides are predicted over a limited time across the H2 event alone, when sea level was near its minimum. We performed an experiment in which the 25-kyr land-plus-ice geometry defined the perimeter of the ocean and present-day values of water-column thickness were used. The Labrador Sea amplitudes were similar to those shown in Fig. 1a, and we conclude that basin geometry, for instance, the existence or absence of ice cover over Hudson Bay,

exerts primary control on Labrador Sea tides.

A numerical study<sup>13</sup> of tidal dissipation over the past 20 kyr also recorded large North Atlantic palaeotides. We have focused on Labrador Sea tides in particular and have shown that they were large over the period between 65 kyr ago and 7 kyr ago. We suggest that these tides preconditioned ice streams and shelves for other forcings, such as climate warming, sea-level rise or ice-stream instabilities<sup>4</sup>, to trigger discrete Heinrich events. Ice-age tides in Europe were comparable to Labrador Sea tides only near the British Isles (Fig. 1a), which may in part explain the greater amount of Canadian material in the ice-rafted debris<sup>3</sup>. The large Labrador Sea palaeotides represent a hitherto unrecognized negative feedback on North American ice-sheet stability, and a potentially important link in our understanding of millennial-scale ice-age climate change.

**Brian K. Arbic\***, **Douglas R. MacAyeal†**,  
**Jerry X. Mitrovica‡**, **Glenn A. Milne§**

\*Program in Atmospheric and Oceanic Sciences,  
Princeton University, PO Box CN710, Sayre Hall,  
Princeton, New Jersey, 08544-0710, USA

e-mail: [arbic@splash.princeton.edu](mailto:arbic@splash.princeton.edu)

†Department of Geophysical Sciences, University of  
Chicago, Chicago, Illinois 60637, USA

‡Department of Physics, University of Toronto,  
Toronto M5S 1A7, Ontario, Canada

§Department of Earth Sciences, University of Durham,  
Science Laboratories, Durham DH1 3LE, UK

1. Rahmstorf, S. *Nature* **419**, 207–214 (2002).
2. Heinrich, H. *Quat. Res.* **29**, 142–152 (1988).
3. Hemming, S. R. *Rev. Geophys.* **42**, RG1005  
doi:10.1029/2003RG000128 (2004).
4. MacAyeal, D. R. *Paleoceanography* **8**, 775–784 (1993).
5. Hulbe, C. L., MacAyeal, D. R., Denton, G. H., Klemm, J. &  
Lowell, T. V. *Paleoceanography* **19**, PA1004  
doi:10.1029/2003PA000890 (2004).
6. Bindshadler, R. A., King, M. A., Alley, R. B., Anandkrishnan, S.  
& Padman, L. *Science* **301**, 1087–1089 (2003).
7. Doake, C. S. M. *et al. Geophys. Res. Lett.* **29**,  
doi:10.1029/1999GL014606 (2002).
8. Lambeck, K., Yokoyama, Y. & Purcell, T. *Quat. Sci. Rev.* **21**,  
343–360 (2002).
9. Platzman, G. W., Curtis, G. A., Hansen, K. S. & Slater, R. D. J.  
*Phys. Oceanogr.* **11**, 579–603 (1981).
10. Arbic, B. K., Garner, S. T., Hallberg, R. W. & Simmons, H. L.  
*Deep-Sea Res. II* (in the press).
11. Milne, G. A., Mitrovica, J. X. & Davis, J. L. *Geophys. J. Int.* **139**,  
464–482 (1999).
12. Egbert, G. D., Bennett, A. F. & Foreman, M. G. G. *J. Geophys.  
Res.* **99**, 24821–24852 (1994).
13. Egbert, G. D., Ray, R. D. & Bills, B. G. *J. Geophys. Res.* **109**,  
C03003, doi:10.1029/2003JC001973 (2004).

Supplementary information accompanies this communication on  
Nature's website.

Competing financial interests: declared none.

# brieF communications arising online

♦ [www.nature.com/bca](http://www.nature.com/bca)

## Palaeoclimatology: Archaean atmosphere and climate

J. F. Kasting (doi:10.1038/nature03166)

## Palaeoclimatology: Archaean palaeosols and Archaean air

N. H. Sleep (doi:10.1038/nature03167)

## Reply

H. Ohmoto & Y. Watanabe (doi:10.1038/nature03168)



## Palaeoclimatology

## Archaeal atmosphere and climate

Arising from: H. Ohmoto, Y. Watanabe & K. Kumazawa *Nature* **429**, 395–399 (2004)

Ohmoto *et al.*<sup>1</sup> argue that carbon dioxide was abundant in the late Archaeal and early Proterozoic atmosphere and that methane was probably scarce, based on a reanalysis of the occurrence of siderite, FeCO<sub>3</sub>, in ancient rocks. Here I consider several factors that may undermine their conclusions.

Rye *et al.*<sup>2</sup> placed an upper limit on atmospheric carbon dioxide partial pressure,  $p\text{CO}_2$ , between 2.2 billion and 2.8 billion years (2.2 and 2.8 Gyr), based on the absence of siderite in palaeosols (soils that are more than 2.5 billion years old). Ohmoto *et al.* discount this result because they argue, quoting my atmospheric models<sup>3,4</sup>, that the oxygen partial pressure ( $p\text{O}_2$ ) was too high to allow siderite to be stable. In doing so, they confuse locally generated  $p\text{O}_2$  values of up to  $10^{-6}$  atm in the immediate vicinity of oxygen oases<sup>4</sup> with globally averaged values of about  $10^{-13}$  atm, and they ignore the much higher partial pressures ( $10^{-5}$  to  $10^{-3}$  atm) of hydrogen and methane in these model atmospheres. If one includes these latter gases in equilibrium groundwater calculations, then the effective  $p\text{O}_2$  is below  $10^{-70}$  atm and siderite is stable, so the Rye *et al.* conclusions stand.

Ohmoto *et al.* argue that the occurrence of siderite in Archaeal/early Proterozoic banded iron formations places a lower limit on atmospheric  $p\text{CO}_2$  at that time, but this argument is not convincing. Banded iron formations are marine sediments that must have been formed below the wave base as otherwise their laminated texture would not have been preserved. Today, the effective  $p\text{CO}_2$  in deep water is about three times higher than that in surface water as a consequence of the decomposition of organic matter<sup>5</sup>. If such gradients in dissolved carbon dioxide existed in the Archaeal era as well, the  $p\text{CO}_2$  estimates derived from this procedure will be systematically high by this same amount.

Two other points are relevant. First, atmospheric methane and carbon dioxide levels need not have been inversely correlated. They are inversely correlated only if one can show that the Archaeal/early Proterozoic climate was cool. This was clearly the case during the Palaeoproterozoic glaciations at around 2.3 Gyr; however, the surface temperature both before and after that time could have been much warmer. A mean temperature of 70 °C has been proposed for 3.3 Gyr ago<sup>6</sup>. If the Archaeal climate was hot, both methane and carbon dioxide could have been abundant. Second, contrary to the assertions by Ohmoto *et al.*, the idea that Archaeal methane levels were high was not proposed initially by Rye *et al.*<sup>2</sup>, nor does it rest exclusively, or even primarily, on

the palaeosol evidence for low  $p\text{CO}_2$ . A methane-rich Archaeal atmosphere was suggested more than 25 years ago<sup>7</sup> and was explored during the 1980s by using both photochemical and climate models<sup>8,9</sup>. If atmospheric oxygen was low<sup>10,11</sup>, and if methanogens are evolutionarily ancient, high methane levels are predicted, regardless of the particulars of Archaeal carbon dioxide levels and climate.

James F. Kasting

Department of Geosciences, Pennsylvania State University, University Park, Pennsylvania 16801, USA  
e-mail: kasting@geosc.psu.edu

doi:10.1038/nature03166

1. Ohmoto, H., Watanabe, Y. & Kumazawa, K. *Nature* **429**, 395–399 (2004).
2. Rye, R., Kuo, P. H. & Holland, H. D. *Nature* **378**, 603–605 (1995).
3. Kasting, J. F. *Precambrian Res.* **34**, 205–229 (1987).
4. Pavlov, A. A., Kasting, J. F. & Brown, L. L. *J. Geophys. Res.* **106**, 23267–23287 (2001).
5. Broecker, W. S. & Peng, T.-H. *Tracers in the Sea* (Lamont Doherty Geol. Obs., Palisades, New York, 1982).
6. Knauth, P. & Lowe, D. R. *Geol. Soc. Am. Bull.* **115**, 566–580 (2003).
7. Walker, J. C. G. *Evolution of the Atmosphere* (Macmillan, New York, 1977).
8. Kasting, J. F., Zahnle, K. J. & Walker, J. C. G. *Precambrian Res.* **20**, 121–148 (1983).
9. Kiehl, J. T. & Dickinson, R. E. *J. Geophys. Res.* **92**, 2991–2998 (1987).
10. Farquhar, J., Bao, H. & Thiemens, M. *Science* **289**, 756–758 (2000).
11. Pavlov, A. A. & Kasting, J. F. *Astrobiology* **2**, 27–41 (2002).

## Palaeoclimatology

## Archaeal palaeosols and Archaeal air

Ferrous carbonate, as the mineral siderite, occurs in Archaeal palaeosols (ancient soils). Ohmoto *et al.*<sup>1</sup> contend that siderite was not in equilibrium with the oxygen in Archaeal air and that its presence in palaeosols provides little constraint on the partial pressure of carbon dioxide in Archaeal air. But their argument is invalid because it fails to distinguish the different behaviours of the trivial component oxygen and the significant component carbon dioxide in the partly closed system of soil waters. The presence or absence of siderite in ancient soils is a valid constraint on the carbon dioxide partial pressure ( $p\text{CO}_2$ ) in ancient atmospheres.

To illustrate the problem using the approach of Ohmoto *et al.*, I use their lowest oxygen partial pressure ( $p\text{O}_2$ ). Solar ultraviolet and lightning would dynamically maintain  $p\text{O}_2$  at  $10^{-13}$  atm (ref. 2). Ohmoto *et al.* correctly state that this is far above the critical  $p\text{O}_2$  for siderite stability, namely  $10^{-65 \pm 5}$  atm. The partial pressure at siderite

stability has only mathematical significance; there is no actual molecule of oxygen in any realizable volume of water at the  $p\text{O}_2$ . Rather, the dissolved oxygen at Archaeal  $p\text{O}_2$  in water, which is about  $10^{-16}$  mol kg<sup>-1</sup>, is too small to affect the chemistry of a soil observably, as illustrated by a simple mass balance. Basalt contains about 10% by mass of ferrous oxide, FeO, so it would take  $2 \times 10^{15}$  kg of oxygen-saturated water to oxidize the ferrous oxide in a kilogram of rock to magnetite, Fe<sub>3</sub>O<sub>4</sub>. With a typical rainfall of 1 m yr<sup>-1</sup>, it would require  $10^{15}$  years of rain to oxidize a metre of section. This conclusion applies as long as oxygen is a trivial component. Even at a partial pressure of  $10^{-6}$  atm, it would take  $10^8$  years to oxidize a metre of section.

Rather, the trivial oxygen dissolved in soil water at  $p\text{O}_2 = 10^{-13}$  atm reacts with the rock, producing an undetectable amount of ferric iron. This leaves a solution that is quantitatively depleted in oxygen and buffered by ferrous silicates, where siderite is stable if there is enough  $p\text{CO}_2$ . At a higher  $p\text{O}_2$  of  $10^{-6}$  atm, trace oxygen in soil water would produce observable ferric iron before it was exhausted, again leaving a buffer with ferrous silicates. Moreover, if the oxidation reaction is kinetically inhibited at trivial concentrations, the presence of oxygen in the air is irrelevant to the presence of siderite.

Norman H. Sleep

Department of Geophysics, Stanford University, Stanford, California 94305, USA

e-mail: norm@pangea.stanford.edu

doi:10.1038/nature03167

1. Ohmoto, H., Watanabe, Y. & Kumazawa, K. *Nature* **429**, 395–399 (2004).
2. Kasting, J. F. *Precambrian Res.* **34**, 205–229 (1987).

Ohmoto *et al.* reply — The idea of a methane-rich Archaeal atmosphere has become popular since Rye *et al.* assumed in their calculation<sup>1</sup> that siderite was absent in pre-2.2-Gyr palaeosols. We have concluded that the absence of siderite in some Archaeal palaeosols does not constrain the atmospheric  $p\text{CO}_2$ , but the presence of much siderite in sedimentary rocks does<sup>2</sup>. Sleep's recognition<sup>3</sup> that siderite occurs in Archaeal palaeosols substantiates our arguments<sup>2</sup>: although siderite should be absent in well aerated soils of all geological ages, it may form in waterlogged soils where  $p\text{O}_2$  became less than about  $10^{-60}$  atm owing to the abundant anaerobic production of H<sub>2</sub>. In fact, we have reported this in a 2.6-Gyr soil profile at Schagen, South Africa<sup>4</sup>: abundant ferric-rich minerals formed while the soil was exposed to air, but ferrous-rich carbonate formed while it was apparently submerged under an anoxic pond.

Cyanobacteria and eukaryotic algae, principal oxygenic photoautotrophs (aerobes), had emerged by about 2.8 Gyr ago<sup>5</sup> and possibly more than 3.7 Gyr ago<sup>6</sup>.

Eukarya require an oxygen-rich environment ( $pO_2 > 0.01$  atm) for sterol synthesis<sup>7</sup>. Waves, currents, tides, rivers, wind and rain spread microbes across the entire globe<sup>8</sup>. Oxygenic phototrophs dominate anoxy-phototrophs in most environments because water is more abundant than other electron donors (such as sulphur and hydrogen sulphide)<sup>9</sup>. The nutrient flux from land to the oceans during the Archaean era was probably higher than today because of more intensive rock weathering<sup>10</sup>. For all these reasons, we suggest that the entire oceanic photic zone became rich in aerobes and dissolved oxygen shortly (probably in much less than 0.1 Myr) after cyanobacteria and eukarya emerged. Considering the flux of oxygen across the air–sea interface, we estimate that the atmospheric  $pO_2$  rose to more than  $10^{-6}$  atm ( $> 1$  p.p.m.) within that time. Therefore, Kasting's<sup>11</sup> arguments for a low- $pO_2$  ( $10^{-13}$  atm) and high- $pH_2$  ( $> 10^{-5}$  atm) atmosphere, calculated from a prebiotic model<sup>12</sup>, and “local oxygen oases”<sup>11,13</sup> do not apply to the post-2.8 Gyr (possibly the post-3.7 Gyr) Earth.

Sleep recognizes that a low- $pO_2$  atmosphere ( $pO_2 < 10^{-6}$  atm) cannot quantitatively explain the oxidation of  $Fe^{2+}$  to  $Fe^{3+}$  observed in pre-2.2-Gyr palaeosols. In fact, from the ferrous–ferric relation in the Schagen palaeosols, we have estimated<sup>4</sup> that the atmospheric  $pO_2$  was already well over  $10^{-3.7}$  atm — much higher than 0.1% of the present atmospheric level (PAL).

Kasting suggests that more than 75% of the dissolved carbon dioxide used by pre-1.8-Gyr massive siderites was derived from the decomposition of organic matter in the oceans and that the remainder was from atmospheric carbon dioxide. If so, the  $\delta^{13}C$  value of these siderites should be very negative (more than  $-20\%$ )<sup>2</sup>. But the most common  $\delta^{13}C$  values are between  $-5\%$  and  $0\%$  (see Fig. 3 in ref. 2), suggesting that more than 80% of the carbon dioxide for the siderites came from the contemporaneous atmosphere. This, in conjunction with the thermodynamic conditions for siderite, constrains the pre-1.8-Gyr atmospheric  $pCO_2$  level to more than  $10^{-1.4 \pm 0.2}$  atm ( $> 100$  PAL) if  $T$  was 25–50 °C and  $SiO_2(aq)$  was less than 100 p.p.m. (ref. 2). If  $T = 70$  °C and  $SiO_2(aq) = 300$  p.p.m. at 3.3 Gyr, then the minimum  $pCO_2$  requirement for the siderites becomes about 2.0 atm; no additional greenhouse gas (methane) was necessary.

It has been suggested<sup>14</sup> that the Archaean atmosphere was methane-rich ( $pCH_4$  of about 1,000 p.p.m., compared with about 1 p.p.m. today), assuming  $T$  was about 300 K and  $pCO_2$  was about 10 PAL at 2.8 Gyr. As methane is a more efficient greenhouse gas than carbon dioxide and our calculations indicate that  $pCO_2 > 100$  PAL, the  $pCH_4$  must have been much less than 100 p.p.m. at 2.8 Gyr. It has also been shown<sup>15</sup> that the  $pCH_4$  value remains below about 3 p.p.m. when  $pO_2 > 1$  p.p.m. Thus, from both the  $pCO_2$ – $pCH_4$  and  $pO_2$ – $pCH_4$  relationships,

we conclude that the Archaean atmosphere was methane-poor. Furthermore, there is no biogeochemical evidence (such as biomarkers) that methanogens emerged earlier than cyanobacteria and created a methane-rich atmosphere.

**Hiroshi Ohmoto, Yumiko Watanabe**

Astrobiology Research Center of the NASA  
Astrobiology Institute & Department of Geosciences,  
Pennsylvania State University, University Park,  
Pennsylvania 16802, USA  
e-mail: ohmoto@geosc.psu.edu  
doi:10.1038/nature03168

1. Rye, R., Kuo, P. H. & Holland, H. D. *Nature* **378**, 603–605 (1995).
2. Ohmoto, H., Watanabe, Y. & Kumazawa, K. *Nature* **429**, 395–399 (2004).
3. Sleep, N. *Nature* doi:10.1038/nature03167 (2004).
4. Watanabe, Y., Stewart, B. W. & Ohmoto, H. *Geochim. Cosmochim. Acta* **68**, 2129–2151 (2004).
5. Brocks, J. J., Buick, R., Logan, G. A. & Summons, R. E. *Geochim. Cosmochim. Acta* **22**, 4321–4335 (2003).
6. Rosing, M. T. & Frei, R. *Earth Planet. Sci. Lett.* **217**, 237–244 (2004).
7. Jahnke, L. & Klein, H. P. *J. Bacteriol.* **155**, 488–492 (1983).
8. Nisbet, E. G. in *Early Precambrian Processes* (eds Coward, M. P. & Ries, A. C.) 27–51 (The Geological Society, London, 1995).
9. Gottschalk, G. in *Bacterial Metabolism* 2nd edn 359 (Springer, New York, 1986).
10. Condie, K. C., Des Marais, D. J. & Abbott, D. *Precambrian Res.* **106**, 239–260 (2001).
11. Kasting, J. F. *Nature* doi:10.1038/nature03166 (2004).
12. Kasting, J. F. *Precambrian Res.* **34**, 205–229 (1987).
13. Kasting, J. F. in *The Proterozoic Biosphere: A Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 1185–1187 (Cambridge Univ. Press, 1992).
14. Pavlov, A. A., Kasting, J. F., Eigenbrode, J. L. & Freeman, K. H. *Geology* **29**, 1003–1006 (2001).
15. Pavlov, A. A. & Kasting, J. F. *Astrobiology* **2**, 27–41 (2002).



# Neanderthals and the modern human colonization of Europe

Paul Mellars

Department of Archaeology, Cambridge University, Downing Street, Cambridge, CB2 3DZ, UK

**The fate of the Neanderthal populations of Europe and western Asia has gripped the popular and scientific imaginations for the past century. Following at least 200,000 years of successful adaptation to the glacial climates of northwestern Eurasia, they disappeared abruptly between 30,000 and 40,000 years ago, to be replaced by populations all but identical to modern humans. Recent research suggests that the roots of this dramatic population replacement can be traced far back to events on another continent, with the appearance of distinctively modern human remains and artefacts in eastern and southern Africa.**

**T**he most significant contributions to these issues over the past decade have come from detailed studies of the DNA structure of present-day human populations in different areas of the world, combined with the gradually accumulating recovery of residual traces of 'ancient' DNA extracted from a number of Neanderthal and early anatomically modern human remains. Studies of both mitochondrial and Y-chromosome DNA patterns in modern world populations (inherited respectively through the female and male lineages) point to the genetic origins of all present-day populations within one limited area of Africa somewhere in the region of 150,000 years before present (yr BP), followed by their dispersal to other regions of the world between about 60,000 and 40,000 yr BP<sup>1–6</sup>. These results are further reinforced by recent discoveries of skeletal remains of anatomically modern populations in different areas. Discoveries at Herto in Ethiopia reported just over a year ago<sup>7</sup> confirm the presence of early forms of anatomically modern humans in Africa by about 160,000 yr BP, whereas the earliest discoveries of distinctively modern populations in both Europe and most parts of Asia can be dated no earlier than 40,000–45,000 yr BP. The one exception is in Israel, where the rich skeletal remains from the Skhul and Qafzeh caves indicate a precocious, and apparently short-lived, incursion of early anatomically modern populations into this region (presumably via the Nile valley) at an early stage in the last glaciation, around 100,000 yr BP<sup>8</sup>.

In Europe, the most dramatic support for these patterns has come from the recovery of a number of relatively well-preserved sequences of mitochondrial DNA from a number of actual skeletal finds of Neanderthals and early anatomically modern humans. Analyses of seven separate Neanderthal specimens (including those from the Neanderthal type-site itself) yielded segments of mitochondrial DNA that are radically different from those of all known present-day populations in either Europe or other parts of the world, and that are equally different from those recovered from five early specimens of anatomically modern populations from European sites<sup>9,10</sup>. The conclusion is clear that there was either very little—if any—interbreeding between the local Neanderthals and the intrusive modern populations in Europe, or that if such interbreeding did take place, all genetic traces of this interbreeding were subsequently eliminated from the European gene pool. The mitochondrial DNA evidence recovered from the Neanderthal specimens further suggests that the initial evolutionary separation of the Neanderthals from the populations which eventually gave rise to genetically modern populations must reach back at least 300,000 yr (ref. 9)—a finding that is in good agreement with the surviving fossil evidence from Africa and Europe<sup>1</sup>. Whether this evidence is sufficient to indicate that the Neanderthals belonged to

an entirely separate biological species from modern humans is at present more controversial<sup>1,2</sup>.

## The archaeological record

One important issue in current research is exactly what patterns of culture and technology were associated with the initial dispersal of anatomically and genetically modern populations across Europe<sup>11,12</sup>. The general assumption in the past has been that this dispersal is represented by the widespread distribution of the 'Aurignacian' technologies, which can be traced continuously from the adjacent areas of the Near East (Israel, Lebanon, Syria and so on) through most areas of eastern and central Europe, to the Atlantic coasts of France and Spain—broadly within the time range from around 40,000 to 35,000 yr BP in conventional (that is, uncalibrated) radiocarbon terms<sup>4,6,8,11–16</sup> (see Fig. 1). Significantly, the Aurignacian period shows an apparently sudden flowering of all the most distinctive features of fully 'modern' (or, in archaeological terms, Upper Palaeolithic) cultural behaviour. Such features include: the first complex and carefully shaped bone, antler and ivory tools; a sudden proliferation of perforated animal teeth, far-travelled marine shells, carefully shaped stone and ivory beads and other forms of personal ornaments; and (at least in sites in central and western Europe) remarkably varied and sophisticated forms of both abstract and figurative art—ranging from engraved outlines of animals, to representations of both male and female sex organs, to the remarkable ivory statuettes of animal and human figures from southern Germany (Fig. 2) and the elaborate cave paintings of the Chauvet cave in southeastern France<sup>14–17</sup>. Collectively, this reflects an explosion in explicitly symbolic behaviour among the Aurignacian populations of Europe and western Asia that is conspicuously lacking from the preceding Middle Palaeolithic Neanderthal communities of the region<sup>18,19</sup>. It is generally agreed that symbolic communication and expression at this level of complexity would be almost inconceivable in the absence of complex language systems and in the absence of brains structured very similarly, if not identically, to our own<sup>20–23</sup>. If we add to this the evidence for the striking uniformity of these Aurignacian technologies across Europe, the sharp break occurring between the earliest Aurignacian and the immediately preceding technologies in the different regions, and the apparent chronological cline in the progressive appearance of this technology from east to west across the continent (Fig. 1), the archaeological evidence alone suggests strongly that the Aurignacian period was that of the initial dispersal of anatomically and behaviourally modern populations across central and western Europe<sup>8,13,15,24</sup>.

## Aurignacian populations

The most frustrating aspect of the current evidence has been the

difficulty of identifying substantial and anatomically distinctive specimens of human skeletal remains in association with Aurignacian technologies<sup>11</sup>. A number of skeletal remains of typically anatomically modern form which were initially attributed to the Aurignacian have recently been shown on the basis of direct radiocarbon dating of the bones themselves to represent intrusive burials into the Aurignacian levels from much later levels—notably those from Velika Pećina in Croatia<sup>25</sup> and those from the remarkable early Aurignacian site of Vogelherd in south Germany (with a range of impressive art objects; see Fig. 2)<sup>12</sup>. The unexpected shock of these discoveries has, perhaps not surprisingly, led some authors to question some of the earlier assumptions about the automatic correlation of the Aurignacian with populations of anatomically modern humans<sup>12,26</sup>.

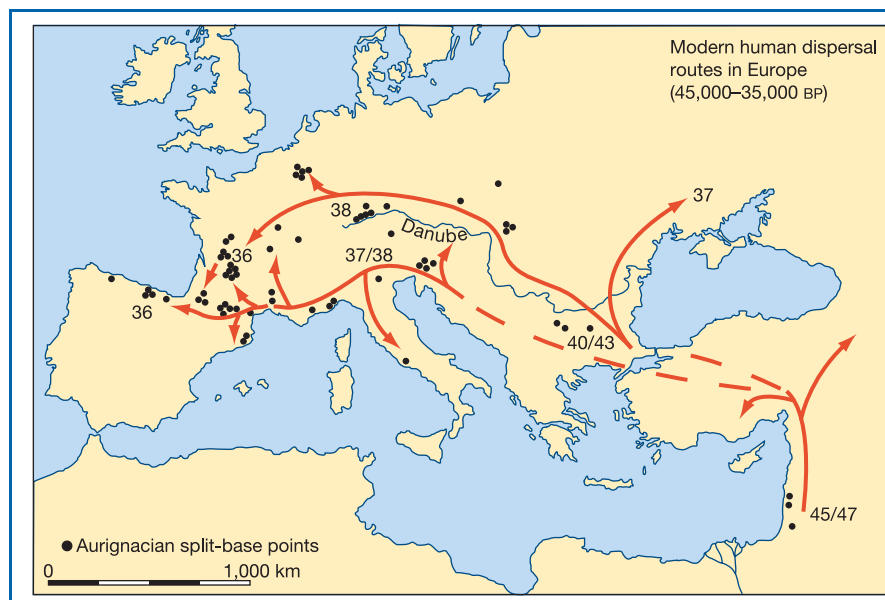
Although understandable, these reactions are nevertheless at best premature and almost certainly unfounded. Even if we implicitly accept the results of all the recent dating evidence, the most we could infer from these results is that the Aurignacians were infuriatingly reluctant to abandon their dead within their main occupation sites—in contrast to the preceding Neanderthals (from whom we now have relatively substantial and well-preserved skeletal remains from at least 20 cave and rock shelter sites), or the populations of the ensuing Gravettian and later stages of the Upper Palaeolithic sequence. To suggest that the scarcity of well-documented skeletal material from the Aurignacian period argues against the association of these populations with anatomically modern humans would be an obvious scientific *non sequitur*. To suggest that this instead favours an association with Neanderthals would be even less defensible.

In reality, the situation is not nearly as bleak as some recent discussions<sup>12,26</sup> have suggested. Despite the elimination of the skeletal material from Vogelherd and Velika Pećina, we now have a range of more fragmentary skeletal remains from at least five or six well-documented contexts in Europe and western Asia which point unmistakably to the presence of diagnostically modern (Cro-Magnon) populations that fall within the time range of the Aurignacian occupation and in several cases are apparently associated directly with Aurignacian archaeological material. The best-dated finds at present are the remains of three typically modern individuals recently reported from the Peștera cu Oase Cave in Romania (directly dated by radiocarbon accelerator measurements to about 35,000 yr BP, although unfortunately not associated directly with archaeological material<sup>26</sup>) and the remains of a complete

juvenile skeleton excavated from the levels immediately underlying the long Aurignacian sequence at Ksar Akil in the Lebanon and dated on the basis of both archaeological evidence and overlying radiocarbon measurements to at least 40,000 yr BP<sup>27,28</sup>. From western Europe we have a fragmentary maxilla from Kents Cavern in Devonshire directly dated to 30,900 ± 900 yr BP<sup>11</sup> and the remains of two characteristically modern mandibles from the site of Les Rois in western France, apparently closely associated with the early Aurignacian levels on the site and dating to around 32,000–35,000 yr BP<sup>29</sup>. Churchill and Smith<sup>11</sup> have suggested that the fragmentary mandible and other remains from the initial pre-Aurignacian (Bachokirian) levels at Bacho Kiro in Bulgaria are probably of anatomically modern form, in this case with radiocarbon dates ranging from 39,000 to 43,000 yr BP. Finally, and perhaps most significantly, the two distinctively anatomically modern crania and other skeletal remains from the site of Mladeč in the Czech Republic have recently been dated on the basis of radiocarbon measurements of associated calcite deposits to around 34,000–35,000 yr BP, and are almost certainly associated with a range of typically Aurignacian bone artefacts<sup>11,30</sup>.

Clearly, there is an urgent need for further direct radiocarbon dating of all these finds to confirm their precise age and associations with typically Aurignacian material. What is beyond dispute at present is that populations that were in most, if not all, respects fully anatomically modern in form were clearly present in several parts of both Europe and the adjacent areas of the Near East entirely within the time range of the Aurignacian period (that is, before around 30,000 yr BP in radiocarbon terms) and well before the appearance of the succeeding Gravettian and later Upper Palaeolithic technologies<sup>14</sup>. This conclusion is reinforced by all the recent studies of both mitochondrial and Y-chromosome DNA patterns in present-day human populations, which point consistently to a dispersal of fully modern (that is, African-derived) patterns of DNA across Europe by at least 35,000 yr BP and probably between 40,000 and 50,000 yr BP<sup>5,6,31</sup>.

And here I should sound a further note of caution. Radiocarbon dating is not without its problems, particularly within the crucial time range of around 30,000–40,000 yr BP under consideration here<sup>32–34</sup>. The problems stem partly from the known fluctuations in the <sup>14</sup>C content of the atmosphere over this time range (which can make measured radiocarbon ages up to 3,000–6,000 years younger than the true, calendar ages of the samples involved<sup>32</sup>) and partly from the potentially serious effects of contamination by



**Figure 1** Apparent dispersal routes of the earliest anatomically and behaviourally modern populations across Europe, as reflected in the archaeological data. The northern route (along the Danube) is represented by the 'classic' Aurignacian technologies, while the southern (Mediterranean) route is represented by the 'proto-Aurignacian' bladelet technologies (Fig. 3)—with their inferred origins in the preceding early Upper Palaeolithic technologies in the Near East and southeastern Europe. Dates (in thousands of years BP) indicate the earliest radiocarbon dates for these technologies in different areas, expressed in thousands of radiocarbon years before present (BP). (These are likely to underestimate the true (calendar) ages of the sites by between 2,000 and 4,000 yr; see ref. 32). Dashed lines indicate uncertain routes.

intrusive recent carbon in the dated samples. For a sample 40,000 years old, contamination by only one per cent of modern carbon would reduce the measured age of the sample by over 6,000 years. This is especially true of dates based on bone or shell samples, which have been shown repeatedly to yield ages that are often several thousand years younger than the true ages, unless extremely rigorous pre-treatment procedures are applied to the samples<sup>33,34</sup>. There is of course no suggestion that the very young ages recently reported for the skeletal remains from Vogelherd and elsewhere can be dismissed in these terms. But in the case of the date of  $27,680 \pm 270$  yr BP reported on a marine shell sample apparently associated with the famous human burials from the Cro-Magnon site in southwest France<sup>35</sup> we cannot exclude the possibility that this represents a true radiocarbon age of  $>30,000$  yr BP, which would put these finds clearly within the Aurignacian time range, as the apparent archaeological associations of the find would suggest<sup>29</sup>. The same caution should be applied to any future applications of direct radiocarbon dating to early Upper Palaeolithic skeletal remains, as well as to those for Neanderthal remains<sup>25</sup>. At least in the case of bone samples, it would be prudent to regard all radiocarbon measurements as minimum estimates for the true age of the samples within the critical range of 30,000–40,000 yr BP (refs 33, 34).

### Patterns of population dispersal

One of the most significant features that has emerged from recent archaeological research is the evidence for an apparently dual pattern of colonization by early anatomically modern populations across Europe, along two different routes. The first route is represented by the distribution of the 'classic' Aurignacian technologies discussed above—that is, those represented at Aurignac itself and

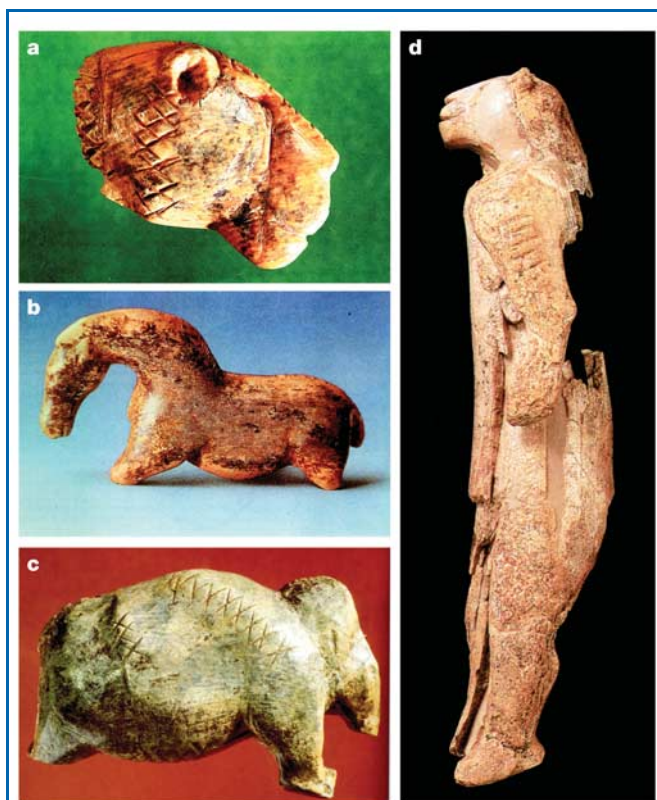
marked by a range of distinctive tool forms, including typical nosed and carinated scraper forms, heavily edge-trimmed Aurignacian blades and perhaps most significantly the highly distinctive split-base bone and antler spear-head forms<sup>13–16</sup> (Fig. 3). As shown in Fig. 1, these technologies are distributed across a broad arc of western, central and southeastern Europe and extend into the immediately adjacent areas of the Near East. At no other point in the Upper Palaeolithic sequence do we observe such a striking similarity in stone and bone technology extending over such a wide diversity of environmental zones. Although the available radiocarbon dates for these technologies show a broadly similar pattern across this region (centred on about 38,000–34,000 yr BP), there are strong indications that the origins of this technology can be identified significantly earlier at sites in southeastern Europe (as at Bacho Kiro and Temnata in Bulgaria) and in the eastern Mediterranean region (as at Ksar Akil in Lebanon) than anywhere in central and western Europe—in both areas extending back to at least 40,000 radiocarbon yr BP<sup>8,13,28,36</sup>. As noted above, the appearance of the new Aurignacian technologies in central and western Europe invariably occurs as an abrupt break with the immediately preceding Neanderthal technologies, strongly supporting their association with new, intrusive populations<sup>13–16</sup>.

The second route of dispersal is distributed mainly along the Mediterranean coast of Europe, extending from at least north-eastern Italy to the Atlantic coast of northern Spain. Although often referred to in the literature as 'archaic' or 'proto' Aurignacian<sup>13,37</sup>, these industries show a very different pattern of technology from that of the classic Aurignacian, dominated mainly by small, carefully shaped bladelets (the Dufour and Font-Yves forms), which probably served as the tips and barbs of composite spear- or arrowheads (Fig. 3). Again, these industries represent a sharp break with the immediately preceding Neanderthal technologies in these areas, and again the most convincing origins for these technologies seem to occur in sites in the Near East (for example, in the lower levels of the Ksar Akil sequence in Lebanon, or at a number of open-air sites such as Boker A in southern Israel) dating back to around 38,000–40,000 yr BP<sup>8,28,36,38</sup>. Both these industries and those of the classic Aurignacian period appear to derive ultimately from the preceding Ahmarian and Emiran technologies of the Near East, clearly represented in the exceptionally long early Upper Palaeolithic sequence at Ksar Akil, and reaching back to at least 45,000–47,000 yr BP<sup>8,28,36</sup>. At Ksar Akil itself these initial Upper Palaeolithic levels are associated with the burial of a typically anatomically modern skeleton<sup>27</sup>. Thus, these levels may well reflect the earliest appearance of fully anatomically modern populations in this region, after their inferred dispersal from Africa shortly before this time.

What is particularly intriguing about these geographical patterns of dispersal of early anatomically modern populations across Europe is their close similarities to the much later dispersal of the earliest agricultural (Neolithic) communities across the continent, between about 10,000 and 6,000 yr BP, that is, comprising a northern route mainly along the Danube valley and a southern route along the Mediterranean coast. To find these close similarities in population dispersal patterns at two widely separated times in European prehistory is one of the most interesting features to have emerged from recent research.

### Neanderthal–modern human interactions

Any model of this kind implies that there must inevitably have been numerous episodes of contact—and therefore potential interaction—between the expanding populations of modern humans and the indigenous Neanderthal populations across Europe. There is insufficient space here to review all of the related discussion that has emerged in the recent literature<sup>17,19,39–41</sup>. One point which now seems clear, however, is that the appearance of a number of apparently modern features of technology among some of the final Neanderthal communities of central and western Europe



**Figure 2** Early Aurignacian carved ivory animal and human figures from sites in southern Germany. **a–c**, Vogelherd Cave; **d**, Hohlenstein–Stadel Cave. The carvings represent the head of a cave lion (**a**), a horse (**b**), a mammoth (**c**) and a male human figure with the head of a cave lion (**d**).



(notably the simple bone tools and a number of grooved or perforated animal-tooth pendants found in the Châtelperronian levels at Arcy-sur-Cure in Central France<sup>17,39,42</sup>) can be shown to coincide closely with the appearance of early Aurignacian populations in the nearby regions of central Europe, and probably with those along the Mediterranean coast<sup>8,14,37,40,43</sup>.

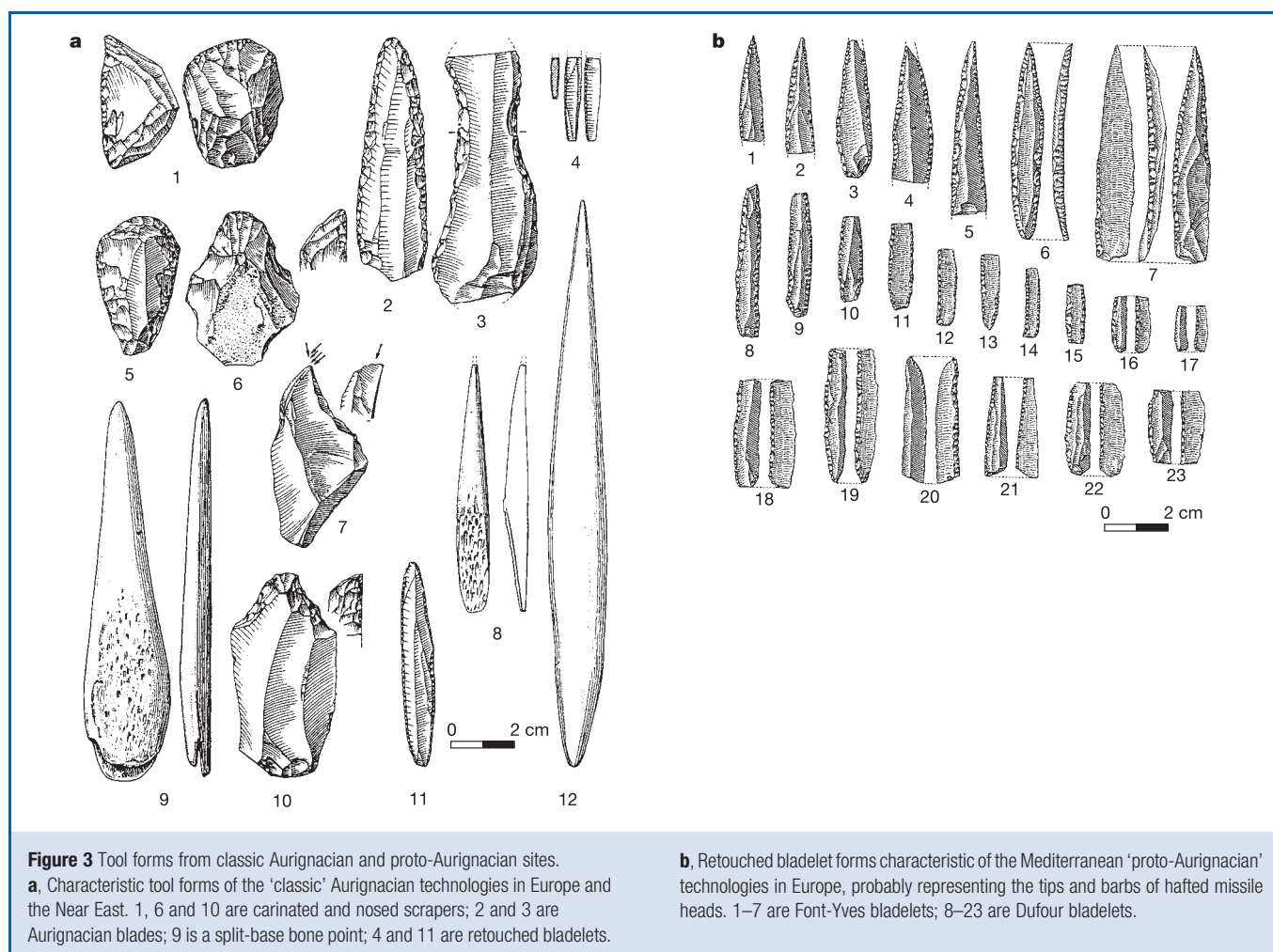
Such patterns of behavioural interaction and technological transfer between the local Neanderthal and intrusive anatomically modern populations are precisely what we would predict on the basis of examples of recent ethnic contact situations<sup>40</sup>, regardless of the respective cultural and cognitive capacities of the two populations. Whether the ability of the final Neanderthals to adopt some of these new patterns of technology can be taken to imply that they had brains effectively identical to those of the incoming modern populations is currently a topic of lively but inconclusive debate<sup>22,23,40,44</sup>. All that can be said is that if the evolutionary trajectories of the Neanderthal and modern populations had been separate for at least 300,000 yr—as all available genetic and anatomical evidence suggests—then the possibility of some divergence in neurological structures over this period cannot be ruled out<sup>45,46</sup>. Equally, the possibility of some small degree of interbreeding between the two populations cannot be excluded on the basis of either the current anatomical or DNA evidence<sup>1,10</sup> and would again seem plausible in anthropological and demographic terms.

However we visualize this situation, the reality is that all traces of distinctively Neanderthal patterns of mitochondrial DNA, as well as the distinctive anatomical features of Neanderthals, disappeared relatively rapidly from European populations<sup>1,2,9,10</sup>. This probably

reflects a straightforward case of direct competition for space and resources between the two populations, in which the demonstrably more complex technology and apparently more complex organization of the anatomically modern populations would have given them a strong competitive advantage over the Neanderthals. Some of the rapid climatic oscillations that have been documented over this time range may also have played a critical part in this demographically competitive situation<sup>47–49</sup>.

### The human revolution?

That the Neanderthals were replaced by populations that had evolved biologically, and no doubt behaviourally, in the very different environments of southern Africa makes the rapid demise of the Neanderthals even more remarkable, and forces us to ask what other cultural or cognitive developments may have made this replacement possible. The rapidly accumulating archaeological evidence for highly symbolic patterns of culture and technology within African populations dating back to at least 70,000 yr BP (marked by the appearance of complex bone technology, multiple-component missile heads, perforated sea-shell ornaments, complex abstract 'artistic' designs and abundant use of red ochre—recently recorded from the Blombos Cave and other sites in southern Africa<sup>50–53</sup>) may provide the critical clue to new patterns of cognition, and probably complex linguistic communication, linked directly with the biological evolution of anatomically and genetically modern populations<sup>1,3</sup>. Perhaps it was the emergence of more complex language and other forms of symbolic communication that gave the crucial adaptive advantage to fully modern



populations and led to their subsequent dispersal across Asia and Europe and the demise of the European Neanderthals. The precise mechanisms and timing of this dramatic population dispersal from southern Africa to the rest of the world remains to be investigated<sup>1,3,4</sup>. □

doi:10.1038/nature03103.

1. Stringer, C. Modern human origins: progress and prospects. *Phil. Trans. R. Soc. Lond. B* **357**, 563–579 (2002).
2. Tattersall, I. in *The Speciation of Modern Homo sapiens* (ed. Crow, T. J.) 49–59 (British Academy, London, 2002).
3. Forster, P. Ice ages and the mitochondrial DNA chronology of human dispersals: a review. *Phil. Trans. R. Soc. Lond. B* **359**, 255–264 (2004).
4. Lahri, M. M. & Foley, R. Towards a theory of modern human origins: geography, demography and diversity in modern human evolution. *Yb. Physical Anthropol.* **41**, 127–176 (1998).
5. Richards, M. *et al.* Tracing European founder lineages in the near Eastern mitochondrial gene pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000).
6. Underhill, P. *et al.* The phylogeography of the Y-chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62 (2001).
7. White, T. D. *et al.* Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* **423**, 742–747 (2003).
8. Bar-Yosef, O. in *The Geography of Neandertals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 107–156 (Peabody Museum, Harvard Univ., Cambridge, Massachusetts, 2000).
9. Krings, M. *et al.* A view of Neanderthal genetic diversity. *Nature Genet.* **26**, 144–146 (2000).
10. Serre, D. *et al.* No evidence of Neanderthal mtDNA contribution to early modern humans. *PLoS Biol.* **2**, 0313–0317 (2004).
11. Churchill, S. E. & Smith, F. H. Makers of the early Aurignacian of Europe. *Yb. Physical Anthropol.* **43**, 61–115 (2000).
12. Conard, N., Grootes, P. M. & Smith, F. H. Unexpectedly recent dates for human remains from Vogelherd. *Nature* **430**, 198–201 (2004).
13. Kozłowski, J. K. & Otte, M. The formation of the Aurignacian in Europe. *J. Anthropol. Res.* **56**, 513–534 (2000).
14. Conard, N. J. & Bolus, M. Radiocarbon dating the appearance of modern humans and timing of cultural innovations in Europe: new results and new challenges. *J. Hum. Evol.* **44**, 331–371 (2003).
15. Mellars, P. A. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. *Phil. Trans. R. Soc. Lond. B* **337**, 225–234 (1992).
16. Mellars, P. A. in *The Peopling of Britain: the Shaping of a Human Landscape* (eds Slack, P. & Ward, R.) 39–67 (Oxford Univ. Press, Oxford, 2001).
17. White, R. Personal ornaments from the Grotte du Renne at Arcy-sur-Cure. *Athena Rev.* **2**, 41–46 (2001).
18. Mellars, P. A. *The Neanderthal Legacy: an Archaeological Perspective from Western Europe* (Princeton Univ. Press, Princeton, New Jersey, 1996).
19. d'Errico, F. The invisible frontier: a multiple species model for the origin of behavioral modernity. *Evol. Anthropol.* **12**, 188–202 (2003).
20. Donald, M. *The Origins of the Modern Mind* (Harvard Univ. Press, Cambridge, Massachusetts, 1991).
21. Pinker, S. *The Language Instinct* (Penguin, London, 1994).
22. Mithen, S. *The Prehistory of the Mind* (Thames & Hudson, London, 1996).
23. Lewis-Williams, D. *The Mind in the Cave* (Thames & Hudson, London, 2002).
24. Davies, W. A very model of a modern human industry: new perspectives on the origins and spread of the Aurignacian in Europe. *Proc. Prehist. Soc.* **67**, 195–217 (2001).
25. Smith, F. H., Trinkaus, E., Pettitt, P. B., Karavanić, I. & Paunović, M. Direct radiocarbon dates for Vindija G1 and Velika Pećina Late Pleistocene hominid remains. *Proc. Natl Acad. Sci. USA* **96**, 12281–12286 (1999).
26. Trinkaus, E. *et al.* An early modern human from Peștera cu Oase, Romania. *Proc. Natl Acad. Sci. USA* **100**, 11231–11236 (2003).
27. Bergman, C. A. & Stringer, C. B. Fifty years after: Egbert, an early Upper Palaeolithic juvenile from Ksar Akil, Lebanon. *Paleorient* **15**, 99–111 (1989).
28. Mellars, P. A. & Tixier, J. Radiocarbon-accelerator dating of Ksar' Aqil (Lebanon) and the chronology

of the Upper Palaeolithic sequence in the Middle East. *Antiquity* **63**, 761–768 (1989).

29. Gambier, D. in *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans* (eds Mellars, P. & Stringer, C.) 194–211 (Edinburgh Univ. Press, Edinburgh, 1989).
30. Svoboda, J., van der Plicht, J. & Kuželka, V. Upper Palaeolithic and Mesolithic human fossils from Moravia and Bohemia (Czech Republic): some new <sup>14</sup>C dates. *Antiquity* **76**, 957–962 (2002).
31. Sherry, S. T. *et al.* Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.* **66**, 761–775 (1994).
32. Shackleton, N. J., Fairbanks, R. G., Chiu, T. & Parrenin, F. Absolute calibration of the Greenland time scale: implications for Antarctic time scales and for  $\Delta^{14}\text{C}$ . *Quat. Sci. Rev.* **23**, 1513–1522 (2004).
33. Aitken, M. J. *Science-based Dating in Archaeology* (Longmans, London, 1990).
34. Bronk Ramsey, C., Higham, T., Bowles, A. & Hedges, R. Improvements to the pretreatment of bone at Oxford. *Radiocarbon* **46**, 155–163 (2004).
35. Henri-Gambier, D. Les fossiles de Cro-Magnon (Les Eyzies-de-Tayac, Dordogne): nouvelles données sur leur position chronologique et leur attribution culturelle. *Bull. Mém. Soc. Anthropol. Paris* **14**, 89–112 (2002).
36. Ohnuma, K. & Bergman, C. A. in *The Emergence of Modern Humans* (ed. Mellars, P.) 91–138 (Edinburgh Univ. Press, Edinburgh, 1990).
37. Le Bon, F. *L'Aurignacien Entre Mer et Océan. Réflexions sur l'unité des phases anciennes de l'Aurignacien dans le Sud de la France* (Société Préhistorique de France, Paris, 2002).
38. Monigal, K. in *More than meets the Eye: Studies on Upper Palaeolithic diversity in the Near East* (eds Goring-Morris, A. N. & Belfer-Cohen, A.) 118–133 (Oxbow books, Oxford, 2003).
39. d'Errico, F., Zilhão, J., Julien, M., Baffier, D. & Pelegrin, J. Neanderthal acculturation in western Europe? A critical review of the evidence and its interpretation. *Curr. Anthropol.* **39**, S1–S44 (1998).
40. Mellars, P. A. The Neanderthal problem continued. *Curr. Anthropol.* **40**, 341–350 (1999).
41. Hublin, J.-J. in *The Geography of Neandertals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 157–182 (Peabody Museum, Harvard Univ., Cambridge, Massachusetts, 2000).
42. Hublin, J.-J., Spoor, F., Braun, M., Zonneveld, F. & Condemi, S. A late Neanderthal associated with Upper Palaeolithic artefacts. *Nature* **381**, 224–226 (1996).
43. Haesaerts, P. & Teyssandier, N. in *The Chronology of the Aurignacian and of the Transitional Complexes* (eds Zilhão, J. & d'Errico, F.) 133–151 (Instituto Português de Arqueologia, Lisbon, 2003).
44. Coolidge, F. L. & Wynn, T. A cognitive and neuropsychological perspective on the Chatelperronian. *J. Anthropol. Res.* **60**, 55–73 (2004).
45. Klein, R. G. Archaeology and the evolution of human behavior. *Evol. Anthropol.* **9**, 7–36 (2000).
46. Enard, W. *et al.* Molecular evolution of FOXP<sub>2</sub>, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
47. Mellars, P. A. in *Neandertals and Modern Humans in Western Asia* (eds Akazawa, T., Aoki, K. & Bar-Yosef, O.) 493–508 (Plenum, New York, 1998).
48. Mellars, P. A. in *Explaining Social Change: Studies in Honour of Colin Renfrew* (eds Cherry, J., Scarre, C. & Shennan, S.) 27–43 (McDonald Institute for Archaeological Research, Cambridge, 2004).
49. Finlayson, C. *Neanderthals, Modern Humans* (Cambridge Univ. Press, Cambridge, 2004).
50. Henshilwood, C. S. *et al.* Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science* **295**, 1278–1280 (2002).
51. Henshilwood, C. S., d'Errico, F., Vanhaeren, M., Van Niekerk, K. & Jacobs, Z. Middle Stone Age shell beads from South Africa. *Science* **304**, 404 (2004).
52. McBrearty, S. & Brooks, A. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.* **39**, 453–563 (2000).
53. Mellars, P. A. in *The Speciation of Modern Homo sapiens* (ed. Crow, T. J.) 31–47 (British Academy, London, 2002).

**Acknowledgements** I am grateful to R. Foley, P. Forster, J.-J. Hublin, J. Kozłowski, C. Stringer and other colleagues for discussions of points raised in the paper, and to D. Kemp for assistance with the illustrations. Research grants were provided by the British Academy and Corpus Christi College, Cambridge.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to P.M. (p.a.mellars@arch.cam.ac.uk).

# Thymosin $\beta$ 4 activates integrin-linked kinase and promotes cardiac cell migration, survival and cardiac repair

Ildiko Bock-Marquette<sup>1,2,5\*</sup>, Ankur Saxena<sup>1,2\*</sup>, Michael D. White<sup>3</sup>, J. Michael DiMaio<sup>3</sup> & Deepak Srivastava<sup>1,2,4</sup>

<sup>1</sup>Department of Pediatrics, <sup>2</sup>Molecular Biology and <sup>3</sup>Cardiovascular and Thoracic Surgery, University of Texas Southwestern Medical Center, 6000 Harry Hines Blvd, Dallas, Texas 75390-9148, USA

<sup>4</sup>Children's Medical Center Dallas, 1935 Motor Street, Dallas, Texas 75390, USA

<sup>5</sup>University of Pécs, Faculty of Medicine, Department of Medical Genetics and Child Development, University of Pécs, H-7624 Pécs, Szigeti u.12., Hungary

\* These authors contributed equally to this work

**Heart disease is a leading cause of death in newborn children and in adults. Efforts to promote cardiac repair through the use of stem cells hold promise but typically involve isolation and introduction of progenitor cells. Here, we show that the G-actin sequestering peptide thymosin  $\beta$ 4 promotes myocardial and endothelial cell migration in the embryonic heart and retains this property in postnatal cardiomyocytes. Survival of embryonic and postnatal cardiomyocytes in culture was also enhanced by thymosin  $\beta$ 4. We found that thymosin  $\beta$ 4 formed a functional complex with PINCH and integrin-linked kinase (ILK), resulting in activation of the survival kinase Akt (also known as protein kinase B). After coronary artery ligation in mice, thymosin  $\beta$ 4 treatment resulted in upregulation of ILK and Akt activity in the heart, enhanced early myocyte survival and improved cardiac function. These findings suggest that thymosin  $\beta$ 4 promotes cardiomyocyte migration, survival and repair and the pathway it regulates may be a new therapeutic target in the setting of acute myocardial damage.**

Coronary artery disease results in acute occlusion of cardiac vessels leading to loss of dependent myocardium. Over thirteen million individuals in the United States alone suffer from coronary artery disease and this condition is one of the leading causes of death in the Western world<sup>1</sup>. Because the heart is incapable of sufficient muscle regeneration, survivors of myocardial infarctions typically develop chronic heart failure. Although more commonly affecting adults, heart disease in children is the leading non-infectious cause of death in the first year of life and often involves abnormalities in cardiac cell specification, migration or survival<sup>2</sup>.

Recent evidence suggests that a population of extracardiac or intracardiac stem cells may contribute to maintenance of the cardiomyocyte population under normal circumstances<sup>3–5</sup>. Although the stem cell population may maintain a delicate balance between cell death and cell renewal, it is insufficient for myocardial repair after acute coronary occlusion. Introduction of isolated stem cells may improve myocardial function<sup>3–5</sup>, but this approach has been controversial<sup>6,7</sup> and requires isolation of autologous stem cells or the use of donor stem cells along with immunosuppression. Technical hurdles of stem cell delivery and differentiation have thus far prevented broad clinical application of cardiac regenerative therapies.

Regulatory pathways involved in cardiac development may have utility in reprogramming cardiomyocytes to aid in cardiac protection or repair<sup>8</sup>. In our studies of genes expressed during cardiac morphogenesis, we found that the 43-amino-acid peptide thymosin  $\beta$ 4 was expressed in the developing heart. Thymosin  $\beta$ 4 has numerous functions, with the most prominent involving sequestration of G-actin monomers and subsequent effects on actin-cytoskeletal organization necessary for cell motility, organogenesis and other cell biological events<sup>9–11</sup>. Recent domain analyses indicate that  $\beta$ -thymosins can affect actin assembly based on their carboxy-terminal affinity for actin<sup>12</sup>. In addition to cell motility, thymosin  $\beta$ 4 may affect transcriptional events by influencing Rho-dependent gene expression or chromatin remodelling events regulated by nuclear actin<sup>13,14</sup>. Although thymosin  $\beta$ 4 promotes skin and corneal wound healing through its effects on cell migration,

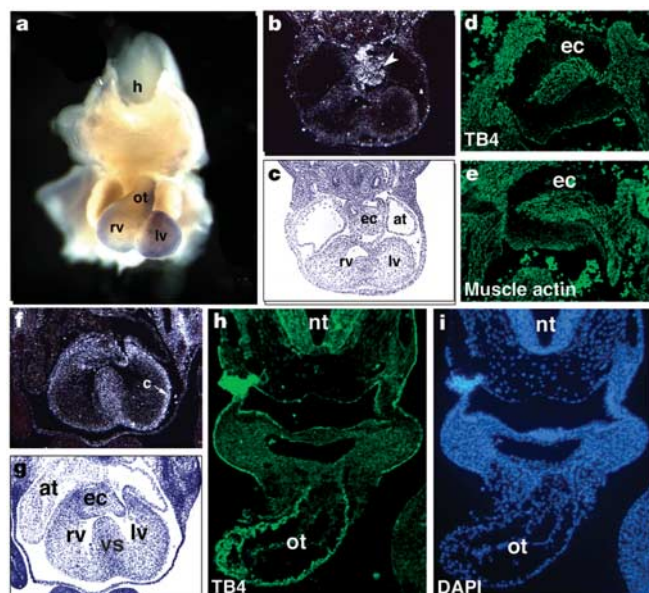
angiogenesis and possibly cell survival<sup>15–17</sup>, the precise molecular mechanism through which it functions and its potential role in solid organ wound healing remain unknown.

Here, we show that thymosin  $\beta$ 4 can stimulate migration of cardiomyocytes and endothelial cells and promotes survival of cardiomyocytes. The LIM domain protein PINCH<sup>18</sup> and ILK<sup>19</sup>, both of which are necessary for cell migration and survival, formed a complex with thymosin  $\beta$ 4 that resulted in phosphorylation of the survival kinase Akt. Inhibition of Akt phosphorylation reversed the effects of thymosin  $\beta$ 4 on cardiac cells. Treatment of adult mice with thymosin  $\beta$ 4 after coronary ligation resulted in increased phosphorylation of Akt in the heart, enhanced early myocyte survival and improved cardiac function. These results indicate that an endogenous protein expressed during cardiogenesis may be re-deployed to protect myocardium in the setting of acute coronary events.

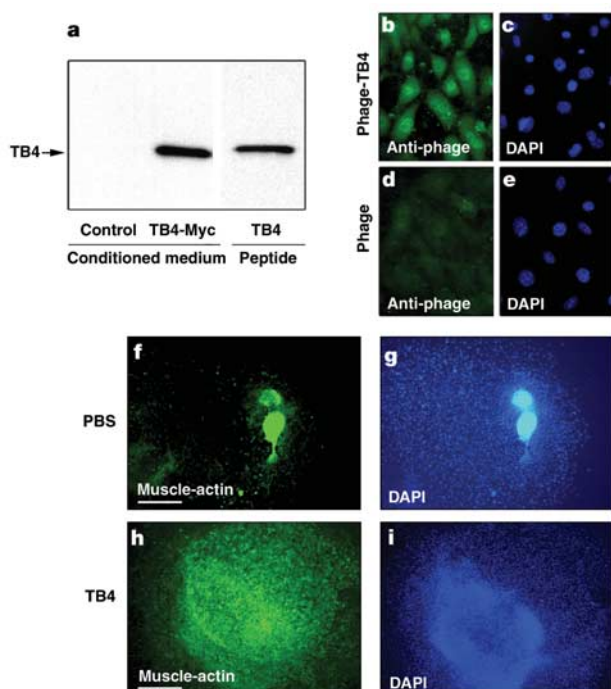
## Developmental expression of thymosin $\beta$ 4

Expression of thymosin  $\beta$ 4 in the developing brain was previously reported<sup>20</sup>, as was expression in the cardiovascular system<sup>21</sup>, although not in significant detail. Whole-mount RNA *in situ* hybridization of embryonic day (E)11.5 mouse embryos revealed thymosin  $\beta$ 4 expression in the left ventricle, outer curvature of the right ventricle and cardiac outflow tract (Fig. 1a). Radioactive *in situ* hybridization indicated that thymosin  $\beta$ 4 transcripts were enriched in the region of cardiac valve precursors known as endocardial cushions (Fig. 1b, c). Cells in this region are derived from endothelial cells that undergo mesenchymal transformation and invade a swelling of extracellular matrix separating the myocardium and endocardium. We found that thymosin- $\beta$ 4-expressing cells in the cushions (Fig. 1d) co-expressed muscle actin (Fig. 1e), suggesting that thymosin  $\beta$ 4 was present in migratory cardiomyocytes known to invade the endocardial cushion<sup>22</sup>. Thymosin  $\beta$ 4 transcripts and protein were also expressed at E9.5–E12.5 in the ventricular septum and the more proliferative region of the myocardium, known as the compact layer, which migrates into the trabecular region as the cells





**Figure 1** Thymosin  $\beta$ 4 is expressed in specific cardiac cell types during development. **a**, Thymosin  $\beta$ 4 mRNA transcripts at E10.5 by whole-mount *in situ* hybridization in frontal view. h, head; lv, left ventricle; ot, outflow tract; rv, right ventricle. **b, c**, Radioactive section *in situ* hybridization at E11.5 in transverse section through heart. Arrowhead indicates endocardial cushion (ec). at, atria. **d, e**, Immunohistochemistry using thymosin  $\beta$ 4 (**d**) and muscle actin (**e**) antibodies focused on cushion cells at E11.5. TB4, thymosin  $\beta$ 4. **f, g**, Expression of thymosin  $\beta$ 4 mRNA at E12.5 in compact layer (c) of ventricles and ventricular septum (vs). Note absence in atria. **h, i**, Thymosin  $\beta$ 4 protein or 4,6-diamidino-2-phenylindole (DAPI) in outflow tract myocardium by immunohistochemistry of E9.5 transverse section. nt, neural tube.



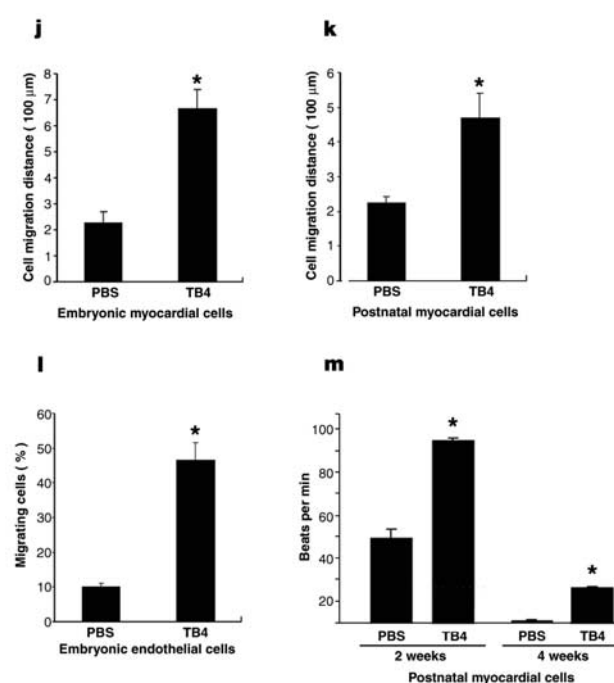
**Figure 2** Thymosin  $\beta$ 4 is secreted and promotes cardiac cell migration and survival. **a**, Western blot of supernatant from thymosin  $\beta$ 4 (TB4) transfected Cos cells using thymosin  $\beta$ 4 antibodies. **b–e**, Immunocytochemistry using anti-phage antibody or DAPI after thymosin- $\beta$ 4-expressing T7 phage (**b, c**) or control phage (**d, e**) administration in the medium of embryonic cardiac explants. **f–i**, Mouse E11.5 cardiac outflow tract explants stained with anti-muscle actin antibody (green) or DAPI (blue) after PBS (**f, g**) or thymosin  $\beta$ 4 (**h, i**) treatment. Scale bars, 500  $\mu$ m. **j, k**, Distance of migrating myocardial cells in

mature (Fig. 1f, g). Finally, outflow tract myocardium that migrates from a secondary heart field also expressed high levels of thymosin  $\beta$ 4 protein<sup>23</sup> (Fig. 1h, i).

### Thymosin $\beta$ 4 induces cardiac cell migration and survival

Although thymosin  $\beta$ 4 is found in the cytosol and nucleus and functions intracellularly<sup>10</sup>, we found that conditioned medium of Cos1 cells transfected with Myc-tagged thymosin  $\beta$ 4 contained thymosin  $\beta$ 4 detectable by western blot (Fig. 2a), consistent with previous reports of thymosin  $\beta$ 4 secretion and presence in wound fluid<sup>17,24,25</sup>. Upon expression of thymosin  $\beta$ 4 on the surface of phage particles added extracellularly to embryonic cardiac explants, we found that an anti-phage antibody coated the cell surface and was ultimately detected intracellularly in the cytosol and nucleus, whereas control phage was not detectable (Fig. 2b–e). Similar observations were made using biotinylated thymosin  $\beta$ 4 (data not shown). These data indicated that secreted thymosin  $\beta$ 4 was internalized into cells, as previously suggested, although the mechanism of cellular entry remains to be determined.

To test the effects of secreted thymosin  $\beta$ 4 on cardiac cell migration, we used an embryonic heart explant system designed to assay cell migration and transformation on a collagen gel<sup>26</sup>. Cardiomyocytes from valve-forming regions secrete signals that induce endocardial cell migration onto collagen, but myocardial cells do not normally migrate in significant numbers (Fig. 2f, g). In contrast, upon addition of thymosin  $\beta$ 4, we observed a large number of spontaneously beating, muscle actin-positive cells that migrated away from the explant (Fig. 2h–j,  $P < 0.0001$ ). No significant difference in cell death or proliferative rate based on TdT-mediated dUTP nick end labelling (TUNEL) assay or phospho-histone H3 immunostaining, respectively, was observed in these cells compared to control cells (data not shown).



E11.5 cardiac outflow tract explants (**j**,  $P < 0.0001$ ) or rat neonatal cardiomyocytes (**k**,  $P < 0.03$ ) with or without thymosin  $\beta$ 4 treatment. **l**, Per cent of embryonic endothelial cells migrating with or without thymosin  $\beta$ 4 ( $P < 0.01$ ). **m**, Beating frequency of rat neonatal cardiomyocytes with or without thymosin  $\beta$ 4 (see Supplementary Figs 1 and 2 for movies;  $P < 0.02$ ). Means and standard deviation bars with 95% confidence limits are shown. Asterisk,  $P < 0.05$ .

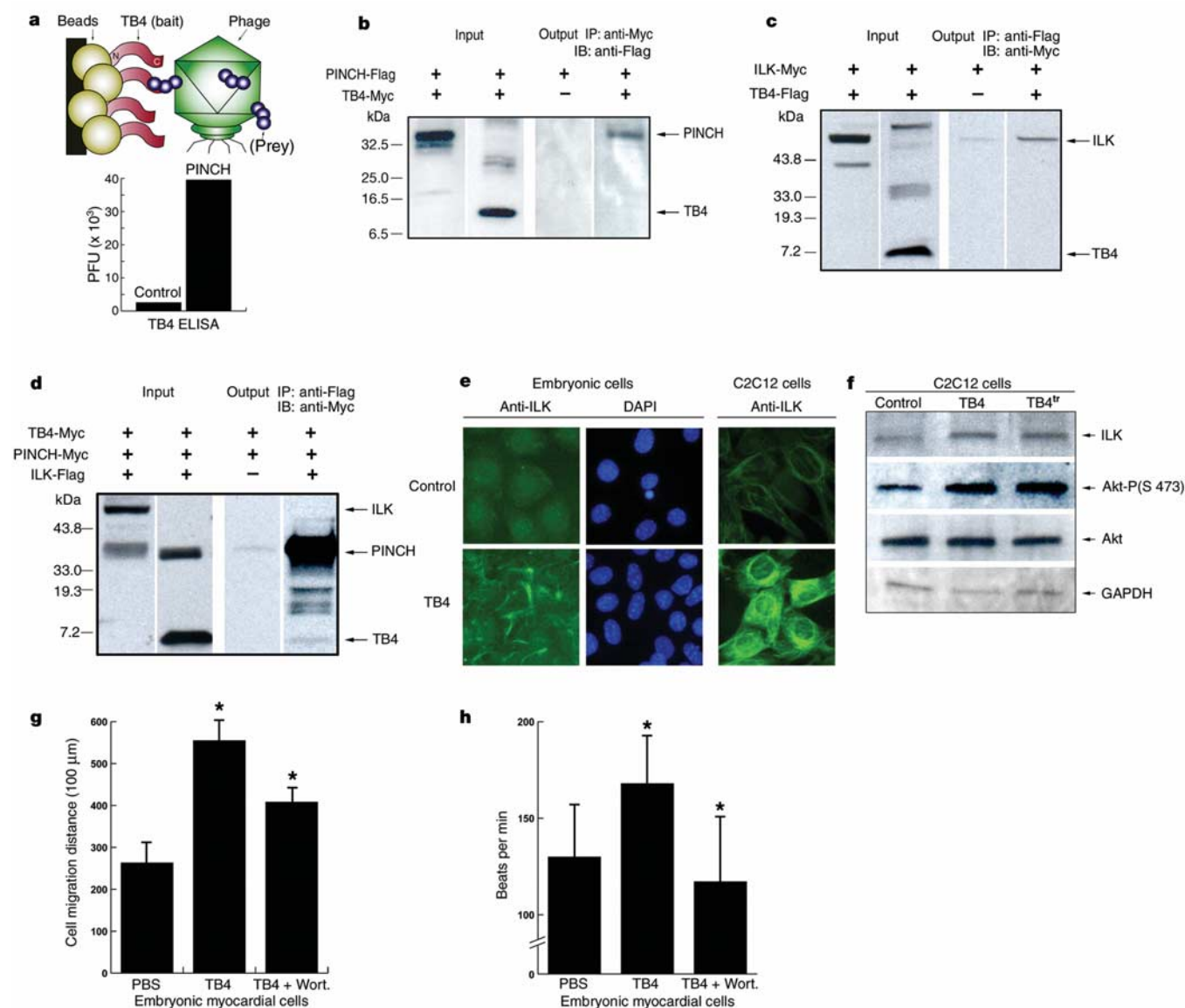
To test the response of postnatal cardiomyocytes, we cultured primary rat neonatal cardiomyocytes on laminin-coated glass and treated the cells with phosphate-buffered saline (PBS) or thymosin  $\beta$ 4. Similar to embryonic cardiomyocytes, the migrational distance of thymosin- $\beta$ 4-treated neonatal cardiomyocytes was significantly increased compared with control (Fig. 2k,  $P < 0.03$ ). In addition to the effects of thymosin  $\beta$ 4 on myocardial cell migration, we observed a similar effect on endothelial migration in the embryonic heart explant assay (Fig. 2l,  $P < 0.01$ ).

Primary culture of neonatal cardiomyocytes typically survives for approximately 1 to 2 weeks, with some cells beating for up to 2 weeks when grown on laminin-coated slides in our laboratory. Surprisingly, neonatal cardiomyocytes survived significantly longer upon exposure to thymosin  $\beta$ 4, with rhythmically contracting

myocytes visible for up to 28 days (Fig. 2m). In addition, the rate of beating was consistently faster in thymosin- $\beta$ 4-treated neonatal cardiomyocytes (95 versus 50 beats per minute,  $P < 0.02$ ), indicating either a change in cell–cell communication or cell metabolism (Fig. 2m; see also Supplementary Figs 1 and 2).

### Thymosin $\beta$ 4 activates ILK and Akt

To investigate the potential mechanisms through which thymosin  $\beta$ 4 might be influencing cell migration and survival events, we searched for thymosin  $\beta$ 4 interacting proteins. The amino terminus of thymosin  $\beta$ 4 was fused with Affi-gel beads resulting in exposure of the C terminus, which allowed identification of previously unknown interacting proteins but prohibited association with actin. We synthesized and screened an E9.5–E12.5 mouse heart T7



**Figure 3** Thymosin  $\beta$ 4 forms a functional complex with PINCH and ILK resulting in phosphorylation of Akt. **a**, Phage display strategy for isolating thymosin  $\beta$ 4 (TB4) interacting proteins, and ELISA confirmation of PINCH interaction. PFU, plaque-forming units. **b**, **c**, Immunoprecipitation (IP) for thymosin  $\beta$ 4 and immunoblot (IB) for PINCH (**b**) or ILK (**c**). **d**, Immunoprecipitation of ILK and immunoblot for PINCH and thymosin  $\beta$ 4. Cell lysate input for each protein is shown along with protein from the immunoprecipitation (output). **e**, Immunocytochemistry with anti-ILK antibody (green) and DAPI (blue) after

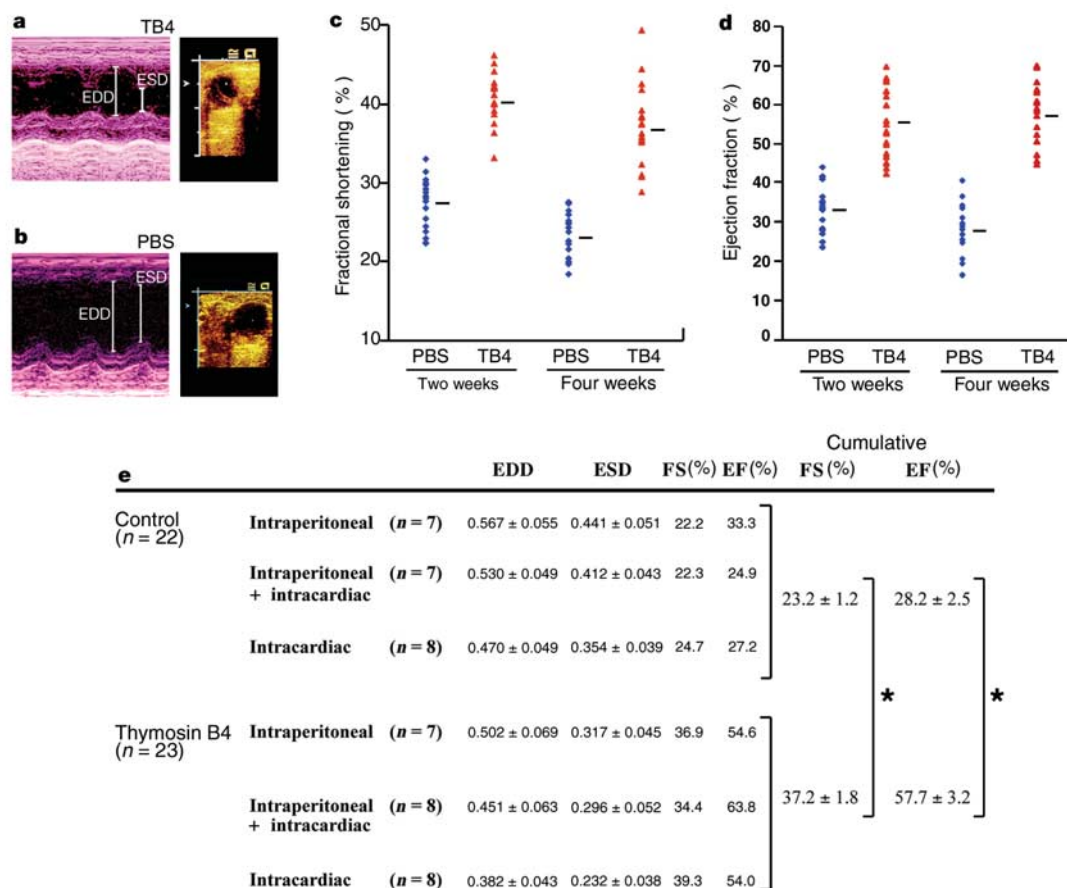
thymosin  $\beta$ 4 treatment of embryonic cardiac explants or C2C12 myoblasts. **f**, Western blot of C2C12 cells treated with thymosin  $\beta$ 4 protein or transfected with thymosin- $\beta$ 4-expressing plasmid (TB4<sup>tr</sup>) using antibodies for ILK, Akt, GAPDH or phospho-specific antibody to Akt-S473. **g**, **h**, Myocardial migration (**g**) or beating frequency (**h**) of E11.5 cardiac explants induced by thymosin  $\beta$ 4 in the presence or absence of wortmannin (Wort.). Bars indicate standard deviations with 95% confidence interval. Asterisk,  $P < 0.05$ .

phage complementary DNA library by phage display, and thymosin- $\beta$ 4-interacting clones were enriched and confirmed by enzyme-linked immunosorbent assay (ELISA, Fig. 3a). PINCH, a LIM domain protein, was most consistently isolated in this screen and interacted with thymosin  $\beta$ 4 in the absence of actin (Fig. 3a). PINCH and ILK interact directly with one another and indirectly with the actin cytoskeleton as part of a larger complex involved in cell–extracellular matrix interactions known as the focal adhesion complex. PINCH and ILK are required for cell motility<sup>18,27</sup> and for cell survival, in part by promoting phosphorylation of the serine-threonine kinase Akt, a central kinase in survival and growth signalling pathways<sup>18,19,27,28</sup>. We transfected plasmids encoding thymosin  $\beta$ 4 with or without PINCH or ILK in cultured cells and found that thymosin  $\beta$ 4 co-precipitated with PINCH or ILK independently (Fig. 3b, c). Moreover, PINCH, ILK and thymosin  $\beta$ 4 consistently immunoprecipitated in a common complex, although the interaction of ILK with thymosin  $\beta$ 4 was weaker than with PINCH (Fig. 3d). The PINCH interaction with thymosin  $\beta$ 4 mapped to the fourth and fifth LIM domains of PINCH, whereas the N-terminal ankyrin domain of ILK was sufficient for thymosin  $\beta$ 4 interaction (data not shown).

Because recruitment of ILK to the focal adhesion complex is important for its activation, we assayed the effects of thymosin  $\beta$ 4 on ILK localization and expression. ILK detection by immunocytochemistry was markedly enhanced around the cell edges after treatment of embryonic heart explants or C2C12 myoblasts with

synthetic thymosin  $\beta$ 4 protein (10 ng per 100  $\mu$ l) or thymosin- $\beta$ 4-expressing plasmid (Fig. 3e). Western analysis indicated a modest increase in ILK protein levels in C2C12 cells, suggesting that the enhanced immunofluorescence may be in part due to altered localization by thymosin  $\beta$ 4 (Fig. 3f). We found that upon thymosin  $\beta$ 4 treatment of C2C12 cells, ILK was functionally activated—evidenced by increased phosphorylation of its known substrate Akt<sup>19</sup> using a phospho-specific antibody to serine 473 of Akt (Fig. 3f)—whereas total Akt protein was unchanged. The similar effects of extracellularly administered thymosin  $\beta$ 4 and transfected thymosin  $\beta$ 4 were consistent with our previous observations of internalization of the peptide, and suggested an intracellular rather than an extracellular role in signalling for thymosin  $\beta$ 4. Because thymosin  $\beta$ 4 sequesters the pool of G-actin monomers, we asked whether the effects on ILK activation were dependent on the role of thymosin  $\beta$ 4 in regulating the balance between polymerized F-actin and monomeric G-actin. We inhibited F-actin polymerization using C3 transferase and also promoted F-actin formation with an activated Rho<sup>29</sup>, but neither intervention affected the ILK levels detected by immunocytochemistry after treatment of COS1 or C2C12 cells with thymosin  $\beta$ 4 (data not shown).

To determine whether activation of ILK was necessary for the observed effects of thymosin  $\beta$ 4, we used a well-described ILK inhibitor, wortmannin, which inhibits ILK's upstream kinase, phosphatidylinositol-3-OH kinase (PI(3)K)<sup>30</sup>. Using myocardial cell migration and beating frequency as assays for thymosin  $\beta$ 4



**Figure 4** Thymosin  $\beta$ 4 treatment after coronary ligation improves myocardial function *in vivo*. **a, b**, Representative echocardiographic M-mode images of left ventricles after coronary ligation with **(a)** or without **(b)** thymosin  $\beta$ 4 (TB4) treatment. Two-dimensional images are shown to the right. **c, d**, Distribution of left ventricular fractional shortening (FS) **(c)** or ejection fraction (EF) **(d)** at 2 and 4 weeks after coronary ligation with (n = 23)

or without (n = 22) thymosin  $\beta$ 4 treatment. Bars indicate means. **e**, Echocardiographic measurements for intraperitoneal, intracardiac or intraperitoneal and intracardiac administration of thymosin  $\beta$ 4 or PBS (Control) at 4 weeks. Means and 95% confidence limits are shown. Asterisk,  $P < 0.0001$ .



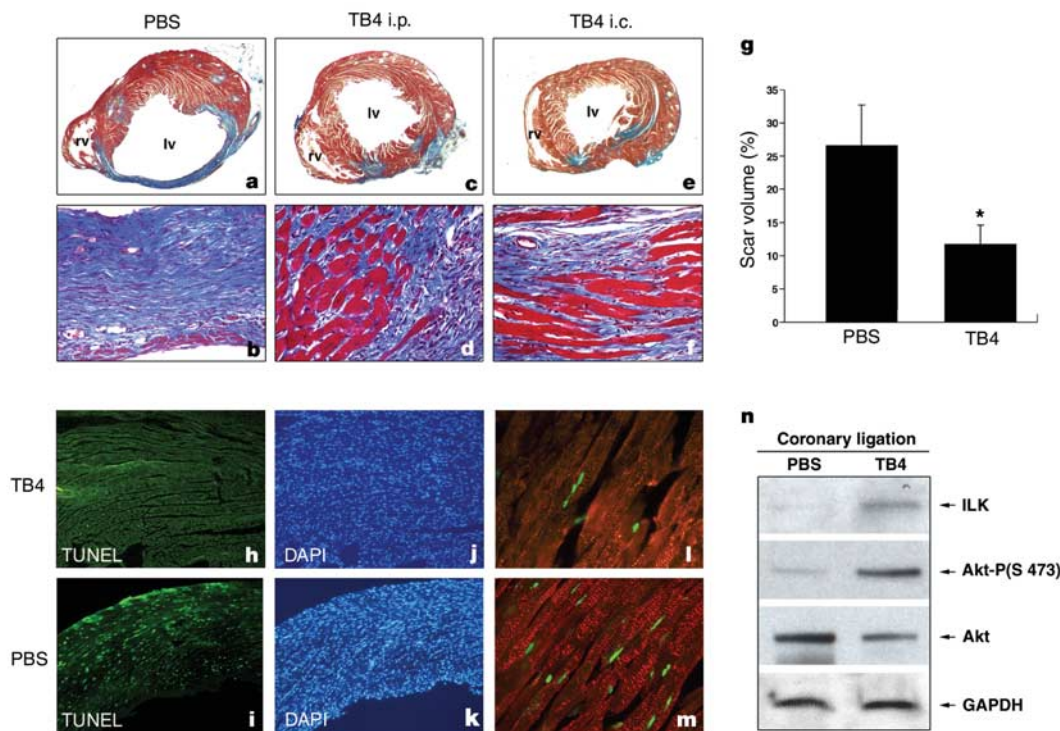
activity, we cultured embryonic heart explants as described above in the presence of thymosin  $\beta 4$  with or without wortmannin. Although inhibiting PI(3)K affects many pathways, we observed a significant reduction in myocardial cell migration and beating frequency upon inhibition of ILK, consistent with ILK mediation of the effects of thymosin  $\beta 4$  (Fig. 3g, h,  $P < 0.05$ ). Together, these results supported a physiologically significant interaction of thymosin  $\beta 4$ –PINCH–ILK within the cell and suggested that this complex may mediate some of the observed effects of thymosin  $\beta 4$  relatively independently of actin polymerization.

### Thymosin $\beta 4$ protects cells after myocardial infarction

Because of the effects of thymosin  $\beta 4$  on cardiac cells *in vitro*, we tested whether thymosin  $\beta 4$  might aid in cardiac repair *in vivo* after myocardial damage. We created myocardial infarctions in 58 adult mice by coronary artery ligation and treated half with systemic, intracardiac, or systemic plus intracardiac thymosin  $\beta 4$  immediately after ligation and the other half with PBS (Fig. 4). All 45 mice that survived 2 weeks later were interrogated for cardiac function by random-blind ultrasonography at 2 and 4 weeks after infarction by multiple measurements of cardiac contraction (Fig. 4a–d). Four weeks after infarction, left ventricles of control mice had a mean fractional shortening of  $23.2 \pm 1.2\%$  ( $n = 22$ , 95% confidence interval); in contrast, mice treated with thymosin  $\beta 4$  had a mean fractional shortening of  $37.2 \pm 1.8\%$  ( $n = 23$ , 95% confidence interval;  $P < 0.0001$ ) (Fig. 4c, e). As a second measure of ventricular function, two-dimensional echocardiographic measurements revealed that the mean fraction of blood ejected from the left ventricle (ejection fraction) in thymosin- $\beta 4$ -treated mice was  $57.7 \pm 3.2\%$  ( $n = 23$ , 95% confidence interval;  $P < 0.0001$ ) com-

pared with a mean of  $28.2 \pm 2.5\%$  ( $n = 22$ , 95% confidence interval) in control mice after coronary ligation (Fig. 4d, e). The greater than 60% or 100% improvement in cardiac fractional shortening or ejection fraction, respectively, suggested a significant improvement with exposure to thymosin  $\beta 4$ , although cardiac function remained depressed compared with sham-operated animals ( $\sim 60\%$  fractional shortening;  $\sim 75\%$  ejection fraction). Finally, the end diastolic dimensions (EDDs) and end systolic dimensions (ESDs) were significantly higher in the control group, indicating that thymosin  $\beta 4$  treatment resulted in decreased cardiac dilation after infarction, consistent with improved function (Fig. 4e). Remarkably, the degree of improvement when thymosin  $\beta 4$  was administered systemically through intraperitoneal injections or only locally within the cardiac infarct was not statistically different, suggesting that the beneficial effects of thymosin  $\beta 4$  probably occurred through a direct effect on cardiac cells rather than through an extracardiac source.

Trichrome stain at three levels of section revealed that the size of scar was reduced in all mice treated with thymosin  $\beta 4$  but was not different between systemic or local delivery of thymosin  $\beta 4$  (Fig. 5a–f), consistent with the echocardiographic data above. Quantification of scar volume using six levels of sections through the left ventricle of a subset of mice demonstrated significant reduction of scar volume in thymosin- $\beta 4$ -treated mice (Fig. 5g,  $P < 0.02$ ). We did not detect significant cardiomyocyte proliferation or death at 3, 6, 11 or 14 days after coronary ligation in PBS or thymosin- $\beta 4$ -treated hearts (data not shown). However, 24 h after ligation we found a marked decrease in cell death by TUNEL assay (green) in thymosin- $\beta 4$ -treated cardiomyocytes (Fig. 5h–k), marked by double-labelling with muscle-actin antibody (red)



**Figure 5** Thymosin  $\beta 4$  promotes survival and alters scar formation after coronary artery ligation in mice. **a–f**, Representative trichrome stain of transverse heart sections at comparable levels 14 days after coronary ligation and PBS (**a**, **b**) or thymosin  $\beta 4$  (TB4) treatment delivered intraperitoneally (i.p.) (**c**, **d**) or intracardiac (i.c.) (**e**, **f**). **b**, **d** and **f** are higher magnifications of **a**, **c** and **e**, respectively. Collagen in scar is indicated in blue and myocytes in red. Images are typical of 20 separate animals. lv, left ventricle; rv, right ventricle. **g**, Estimated scar volume of hearts after coronary ligation and PBS or thymosin

$\beta 4$  treatment. Bars indicate standard deviation at 95% confidence limits. Asterisk,  $P < 0.02$ . **h**, **i**, TUNEL-positive cells (bright green) 24 h after coronary ligation and thymosin  $\beta 4$  or PBS treatment. **j**, **k**, DAPI stain of **h**, **i**. **l**, **m**, Higher magnification of TUNEL-positive nuclei (green) double-labelled with anti-muscle actin antibody (red striations) to mark cardiomyocytes. **n**, Western blot on heart lysates after coronary ligation and treatment with PBS or thymosin  $\beta 4$ .

(Fig. 5l, m). TUNEL-positive cells that were also myocytes were rare in the thymosin  $\beta 4$  group but abundant in the control hearts. Consistent with this observation, we found that the left ventricle fractional shortening 3 days after infarction was  $39.2 \pm 2.3\%$  ( $n = 4$ , 95% confidence interval) with intracardiac thymosin  $\beta 4$  treatment compared with  $28.8 \pm 2.3\%$  ( $n = 4$ , 95% confidence interval) in controls ( $P < 0.02$ ); ejection fraction was  $64.2 \pm 6.7\%$  or  $44.7 \pm 8.4\%$ , respectively ( $P < 0.02$ ), suggesting early protection by thymosin  $\beta 4$ . Finally, we failed to detect any differences in the number of c-kit, Sca-1 or Abcg2 positive cardiomyocytes between treated and untreated hearts, and the cell volume of cardiomyocytes in thymosin- $\beta 4$ -treated animals was similar to mature myocytes, suggesting that the thymosin- $\beta 4$ -induced improvement was unlikely to be influenced by recruitment of known stem cells into the cardiac lineage (data not shown). Thus, the decreased scar volume and preserved function of thymosin- $\beta 4$ -treated mice were probably due to early preservation of myocardium after infarction through the effects of thymosin  $\beta 4$  on survival of cardiomyocytes.

Similar to cultured cells, the level of ILK protein was increased in heart lysates of mice treated with thymosin  $\beta 4$  after coronary ligation compared with PBS-treated mice (Fig. 5n). Correspondingly, phospho-specific antibodies to Akt-S473 revealed an elevation in the amount of phosphorylated Akt-S473 in mice treated with thymosin  $\beta 4$  (Fig. 5n), consistent with the effects of thymosin  $\beta 4$  on ILK described earlier (Fig. 3e, f). These observations *in vivo* were consistent with the effects of thymosin  $\beta 4$  on cell migration and survival demonstrated *in vitro*, and suggest that activation of ILK and subsequent stimulation of Akt may in part explain the enhanced cardiomyocyte survival induced by thymosin  $\beta 4$ , although it is unlikely that a single mechanism is responsible for the full repertoire of thymosin  $\beta 4$ 's cellular effects.

## Discussion

The evidence presented here suggests that thymosin  $\beta 4$ , a protein involved in cell migration and survival during cardiac morphogenesis, may be re-deployed to minimize cardiomyocyte loss after cardiac infarction. Given the known roles of PINCH, ILK and Akt, our data are consistent with this complex having a central role in the effects of thymosin  $\beta 4$  on cell motility, survival and cardiac repair. The ability of thymosin  $\beta 4$  to prevent cell death within 24 h after coronary ligation probably leads to the decreased scar volume and improved ventricular function observed in mice. Although thymosin  $\beta 4$  activation of ILK is likely to have many cellular effects, the activation of Akt may be the dominant mechanism through which thymosin  $\beta 4$  promotes cell survival. This is consistent with Akt's proposed effect on cardiac repair when overexpressed in mouse marrow-derived stem cells administered after cardiac injury<sup>31</sup>, although this probably occurs in a non-cell-autonomous fashion. Whereas thymosin  $\beta 4$  can augment an organism's ability to heal surface wound and stimulate angiogenesis<sup>16,17,32</sup>, the work presented here is the first demonstration of thymosin  $\beta 4$ 's efficacy in healing of a solid organ, and reveals a new mechanism through which thymosin  $\beta 4$  affects cellular functions. Whether thymosin  $\beta 4$  directly affects stabilization of ILK or transcription of ILK through actin-dependent regulation of transcription factors, and which cell types are affected by these or other pathways, remain to be determined.

The early effect of thymosin  $\beta 4$  in protecting the heart from cell death was reminiscent of myocytes that are able to survive hypoxic insult by "hibernating"<sup>33</sup>. Although the mechanisms underlying hibernating myocardium are unclear, alterations in metabolism and energy usage seem to promote survival of cells<sup>33</sup>. Future studies will determine whether thymosin  $\beta 4$  alters cellular properties in a manner similar to hibernating myocardium, possibly allowing time for endothelial cell migration and new blood vessel formation. Given the findings here, thymosin  $\beta 4$ , or the discovery of small

molecules that mimic its function, may prove useful for protecting patients from cardiac injury and therefore warrant further pre-clinical investigation. □

## Methods

### RNA *in situ* hybridization

Whole-mount or section RNA *in situ* hybridization of E9.5–E12.5 mouse embryos was performed with digoxigenin-labelled or <sup>35</sup>S-labelled antisense riboprobes synthesized from the 3' untranslated region of mouse thymosin  $\beta 4$  cDNA that did not share homology with the closely related transcript of thymosin  $\beta 10$ , as previously described<sup>34</sup>.

### Immunohistochemistry

Embryonic or adult cardiac tissue was embedded in paraffin and sections used for immunohistochemistry. Embryonic heart sections were incubated with anti-thymosin  $\beta 4$  (a gift of H. Yin) that does not recognize thymosin  $\beta 10$  (ref. 35). Adult hearts were sectioned at ten equivalent levels from the base of the heart to the apex. Serial sections were used for trichrome sections and reaction with muscle actin, c-kit, Sca-1, Abcg2 and BrdU antibodies and for TUNEL assay (Intergen Company S7111).

### Collagen gel migration assay

Outflow tract was dissected from E11.5 wild-type mouse embryos and placed on collagen matrices as previously described<sup>26</sup>. After 10 h of attachment explants were incubated in 30 ng per 300  $\mu$ l thymosin  $\beta 4$  in PBS, PBS alone or thymosin  $\beta 4$  and 100 nM wortmannin. Cultures were carried out for 3–9 days at 37 °C 5% CO<sub>2</sub> and fixed in 4% paraformaldehyde in PBS for 10 min at room temperature. Cells were counted for quantification of migration and distance using at least three separate explants under each condition for endothelial migration and eight separate explants for myocardial migration.

### Immunocytochemistry on collagen gel explants

Paraformaldehyde-fixed explants were permeabilized for 10 min at room temperature with Permeabilize solution (10 mM PIPES pH 6.8, 50 mM NaCl, 0.5% Triton X-100, 300 mM sucrose, 3 mM MgCl<sub>2</sub>) and rinsed with PBS twice for 5 min each at room temperature. After a series of blocking and rinsing steps, detection antibodies were used and explants rinsed and incubated with equilibration buffer (Anti-Fade kit) for 10 min at room temperature. Explants were scooped to a glass microscope slide, covered and examined by fluorescein microscopy. TUNEL assay was performed using ApopTag plus fluorescein *in situ* apoptosis detection kit (Intergen Company S7111) as recommended.

### Embryonic T7 phage display cDNA library and phage biopanning

Equal amounts of messenger RNA were isolated and purified from E9.5–E12.5 mouse embryonic hearts by using Straight A's mRNA Isolation System (Novagen). cDNA was synthesized by using T7Select10-3 OrientExpress cDNA Random Primer Cloning System (Novagen). The vector T7Select10-3 was used to display random-primed cDNA at the C terminus of 5–15 phage 10B coat protein molecules. 10<sup>9</sup> plaque-forming units of the T7 phage embryonic heart library (100 × of the complexity) in 500  $\mu$ l of PBST was applied to a column of Affi-gel bound to thymosin  $\beta 4$  to achieve low-stringency biopanning to identify thymosin  $\beta 4$  interacting partners. See Supplementary Methods for details of phage packaging, phage biopanning and ELISA confirmation.

### Co-immunoprecipitation

Cos1 and 10T1/2 cells were transfected with thymosin  $\beta 4$ , PINCH and/or ILK and lysates precipitated with antibodies to each as previously described<sup>36</sup>. Western blots were performed using anti-ILK polyclonal antibody (Santa Cruz), anti-thymosin  $\beta 4$  polyclonal antibody<sup>35</sup> (gift of H. Yin) and anti-Myc or anti-Flag antibody against tagged versions of PINCH.

### Animals and surgical procedures

Myocardial infarction was produced in 58 male C57BL/6J mice at 16 weeks of age (25–30 g) by ligation of the left anterior descending coronary artery as previously described<sup>37</sup>. All animal protocols were reviewed and approved by the University of Texas Southwestern Medical Center Institutional Animal Care Advisory Committee and were in compliance with the rules governing animal use as published by the NIH. Twenty-nine of the ligated mice received thymosin  $\beta 4$  treatment immediately after ligation and the remaining 29 received PBS injections. Treatment was given intracardiac with thymosin  $\beta 4$  (400 ng in 10  $\mu$ l collagen) or with 10  $\mu$ l of collagen; intraperitoneally with thymosin  $\beta 4$  (150  $\mu$ g in 300  $\mu$ l PBS) or with 300  $\mu$ l of PBS; or by both intracardiac and intraperitoneal injections. Intraperitoneal injections were given every 3 days until mice were killed. Doses were based on previous studies of thymosin  $\beta 4$  biodistribution<sup>38</sup>. Hearts were removed, weighed and fixed for histological sectioning. Additional mice were operated on in a similar fashion for studies 0.5, 1, 3, 6 and 11 days after ligation.

### Analysis of cardiac function by echocardiography

Echocardiograms to assess systolic function were performed using M-mode and two-dimensional measurements as described previously<sup>37</sup>. The measurements represented the average of six selected cardiac cycles from at least two separate scans performed in random-blind fashion with papillary muscles used as a point of reference for consistency in level of scan. End diastole was defined as the maximal left ventricle diastolic dimension and end systole was defined as the peak of posterior wall motion. Single outliers in each group were omitted for statistical analysis. Fractional shortening (FS), a surrogate of systolic function, was calculated from left ventricle dimensions as follows:

$FS = ((EDD - ESD)/EDD) \times 100\%$ . Ejection fraction (EF) was calculated from two-dimensional images.

## Calculation of scar volume

Scar volume was calculated using six sections through the heart of each mouse using Openlab 3.03 software (Improvision) similar to that previously described<sup>6</sup>. Per cent area of collagen deposition was measured on each section in a blinded fashion and averaged for each mouse.

## Statistical analyses

Statistical calculations were performed using a standard *t*-test of variables with 95% confidence intervals.

Received 15 July; accepted 10 September 2004; doi:10.1038/nature03000.

1. American Heart Association, *Heart Disease and Stroke Statistics—2004 update* 11–14 (American Heart Association, Dallas, Texas, 2004).
2. Hoffman, J. I. E. & Kaplan, S. The incidence of congenital heart disease. *J. Am. Coll. Cardiol.* **39**, 1890–1900 (2002).
3. Orlic, D. *et al.* Bone marrow cells regenerate infarcted myocardium. *Nature* **401**, 701–705 (2001).
4. Beltrami, A. P. *et al.* Adult cardiac stem cells are multipotent and support myocardial regeneration. *Cell* **114**, 763–776 (2003).
5. Anversa, P. & Nadal-Ginard, B. Myocyte renewal and ventricular remodelling. *Nature* **415**, 240–243 (2002).
6. Balsam, L. B. *et al.* Haematopoietic stem cells adopt mature haematopoietic fates in ischaemic myocardium. *Nature* **428**, 668–673 (2004).
7. Murry, C. E. *et al.* Haematopoietic stem cells do not transdifferentiate into cardiac myocytes in myocardial infarcts. *Nature* **428**, 664–668 (2004).
8. Srivastava, D. & Olson, E. N. A genetic blueprint for cardiac development: Implications for human heart disease. *Nature* **407**, 221–226 (2000).
9. Safer, D., Elzinga, M. & Nachmias, V. T. Thymosin  $\beta$ 4 and Fx, an actin-sequestering peptide, are indistinguishable. *J. Biol. Chem.* **266**, 4029–4032 (1991).
10. Huff, T., Muller, C. S., Otto, A. M., Netzker, R. & Hannappel, E. Beta-Thymosins, small acidic peptides with multiple functions. *Int. J. Biochem. Cell Biol.* **33**, 205–220 (2001).
11. Sun, H. Q., Kwiatkowska, K. & Yin, H. L.  $\beta$ -Thymosins are not simple actin monomer buffering proteins. Insights from overexpression studies. *J. Biol. Chem.* **271**, 9223–9230 (1996).
12. Hertzog, M. *et al.* The  $\beta$ -thymosin/WH2 domain; structural basis for the switch from inhibition to promotion of actin assembly. *Cell* **117**, 611–623 (2004).
13. Marinissen, M. J. *et al.* Small GTP-binding protein RhoA regulates c-jun by a ROCK-JNK signaling axis. *Mol. Cell* **14**, 29–41 (2004).
14. Olave, I. A., Reck-Peterson, S. L. & Crabtree, G. R. Nuclear actin and actin-related protein in chromatin remodeling. *Annu. Rev. Biochem.* **71**, 755–781 (2002).
15. Malinda, K. M. *et al.* Thymosin  $\beta$ 4 accelerates wound healing. *J. Invest. Dermatol.* **113**, 364–368 (1999).
16. Sosne, G. *et al.* Thymosin  $\beta$ 4 promotes corneal wound healing and decreases inflammation *in vivo* following alkali injury. *Exp. Eye Res.* **74**, 293–299 (2002).
17. Grant, D. S. *et al.* Thymosin  $\beta$ 4 enhances endothelial cell differentiation and angiogenesis. *Angiogenesis* **3**, 125–135 (1999).
18. Fukuda, T., Chen, K., Shi, X. & Wu, C. PINCH-1 is an obligate partner of integrin-linked kinase (ILK) functioning in cell shape modulation, motility, and survival. *J. Biol. Chem.* **278**, 51324–51333 (2003).
19. Troussard, A. A. *et al.* Conditional knock-out of integrin-linked kinase demonstrates an essential role in protein kinase B/Akt activation. *J. Biol. Chem.* **278**, 22374–22378 (2003).
20. Lin, S. C. & Morrison-Bogorad, M. Developmental expression of mRNAs encoding TB4 and TB10 in rat brain and other tissues. *J. Mol. Neurosci.* **2**, 35–44 (1990).
21. Gomez-Marquez, J., Franco del Amo, F., Carpintero, P. & Anadon, R. High levels of mouse thymosin

- $\beta$ 4 mRNA in differentiating P19 embryonic cells and during development of cardiovascular tissues. *Biochim. Biophys. Acta* **1306**, 187–193 (1996).
22. Van den Hoff, M. J. *et al.* Myocardialization of the cardiac outflow tract. *Dev. Biol.* **212**, 477–490 (1999).
23. Kelly, R. G. & Buckingham, M. E. The anterior heart-forming field: voyage to the arterial pole of the heart. *Trends Genet.* **18**, 210–216 (2002).
24. Frohm, M. *et al.* Biochemical and antibacterial analysis of human wound and blister fluid. *Eur. J. Biochem.* **237**, 86–92 (1996).
25. Huang, W. Q. & Wang, Q. R. Bone marrow endothelial cells secrete thymosin  $\beta$ 4 and AcSDKP. *Exp. Hematol.* **29**, 12–18 (2001).
26. Runyan, R. B. & Markwald, R. R. Invasion of mesenchyme into three-dimensional collagen gels: a regional and temporal analysis of interaction in embryonic heart tissue. *Dev. Biol.* **95**, 108–114 (1983).
27. Zhang, Y. *et al.* Assembly of the PINCH-ILK-CH-ILKBP complex precedes and is essential for localization of each component to cell-matrix adhesion sites. *J. Cell Sci.* **115**, 4777–4786 (2002).
28. Brazil, D. P., Park, J. & Hemmings, B. A. PKB binding proteins. Getting in on the Akt. *Cell* **111**, 293–303 (2002).
29. Arai, A., Spencer, J. A. & Olson, E. N. STARS, a striated muscle activator of Rho signaling and serum response factor-dependent transcription. *J. Biol. Chem.* **277**, 24453–24459 (2002).
30. Delcommenne, M. *et al.* Phosphoinositide-3-OH kinase-dependent regulation of glycogen synthase kinase 3 and protein kinase B/AKT by the integrin-linked kinase. *Proc. Natl Acad. Sci. USA* **95**, 11211–11216 (1998).
31. Mangi, A. A. *et al.* Mesenchymal stem cells modified with Akt prevent remodeling and restore performance of infarcted hearts. *Nature Med.* **9**, 1195–1201 (2003).
32. Philp, D. *et al.* Thymosin  $\beta$ 4 and a synthetic peptide containing its actin-binding domain promote dermal wound repair in db/db diabetic mice and in aged mice. *Wound Rep. Reg.* **11**, 19–24 (2003).
33. Depre, C. *et al.* Program of cell survival underlying human and experimental hibernating myocardium. *Circ Res.* **95**, 433–440 (2004).
34. Yamagishi, H. *et al.* *Tbx1* is regulated by tissue-specific forkhead proteins through a common sonic hedgehog-responsive enhancer. *Genes Dev.* **17**, 269–281 (2003).
35. Yu, F. X., Lin, S. C., Morrison-Bogorad, M., Atkinson, M. A. & Yin, H. L. Thymosin beta 10 and thymosin beta 4 are both actin monomer sequestering proteins. *J. Biol. Chem.* **268**, 502–509 (1993).
36. Garg, V. *et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
37. Garner, L. B. *et al.* Macrophage migration inhibitory factor is a cardiac-derived myocardial depressant factor. *Am. J. Physiol. Heart Circ. Physiol.* **285**, 500–509 (2003).
38. Mora, C. A., Baumann, C. A., Paino, J. E., Goldstein, A. L. & Badamchian, M. Biodistribution of synthetic thymosin  $\beta$ 4 in the serum, urine, and major organs of mice. *Int. J. Immunopharmacol.* **19**, 1–8 (1997).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** The authors wish to thank A. L. Goldstein for his advice and RegeneRx Biopharmaceuticals Inc. for providing the synthetic thymosin  $\beta$ 4 protein; J. Richardson and the histopathology core for histological support; G. A. Adams for technical assistance; E. N. Olson and members of the Srivastava laboratory for discussions and critical review; and S. Johnson and J. E. Marquette for graphical help and suggestions. D.S. was supported by grants from the NHLBI/NIH, March of Dimes Birth Defects Foundation, American Heart Association and Donald W. Reynolds Clinical Cardiovascular Research Center.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to D.S. (Deepak.Srivastava@UTSouthwestern.edu).



# Human DNA ligase I completely encircles and partially unwinds nicked DNA

John M. Pascal<sup>1</sup>, Patrick J. O'Brien<sup>1\*</sup>, Alan E. Tomkinson<sup>2</sup> & Tom Ellenberger<sup>1</sup>

<sup>1</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA

<sup>2</sup>Radiation Oncology Research Laboratory and Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA

\* Present address: Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA

**The end-joining reaction catalysed by DNA ligases is required by all organisms and serves as the ultimate step of DNA replication, repair and recombination processes. One of three well characterized mammalian DNA ligases, DNA ligase I, joins Okazaki fragments during DNA replication. Here we report the crystal structure of human DNA ligase I (residues 233 to 919) in complex with a nicked, 5' adenylated DNA intermediate. The structure shows that the enzyme redirects the path of the double helix to expose the nick termini for the strand-joining reaction. It also reveals a unique feature of mammalian ligases: a DNA-binding domain that allows ligase I to encircle its DNA substrate, stabilizes the DNA in a distorted structure, and positions the catalytic core on the nick. Similarities in the toroidal shape and dimensions of DNA ligase I and the proliferating cell nuclear antigen sliding clamp are suggestive of an extensive protein–protein interface that may coordinate the joining of Okazaki fragments.**

Shortly after DNA ligases were discovered in the late 1960s, it was found that enzymatic ligation consists of three chemical reactions coupled through two covalent intermediates: an enzyme–adenylate and a DNA–adenylate (Fig. 1a)<sup>1</sup>. Sequence comparisons<sup>2</sup> and extensive mutational analysis<sup>3–7</sup> have since identified many key residues involved in the DNA ligation reaction. X-ray structures of DNA ligases<sup>8–11</sup> have revealed both general structural features<sup>12,13</sup> and specific molecular insights regarding the first step of the reaction, which can occur in the absence of DNA. In contrast, considerably less is known about the subsequent steps that involve the DNA substrate. Here human DNA ligase I (Lig1) was crystallized in complex with 5'-adenylated DNA (AppDNA) after reacting a nicked DNA substrate with an amino-terminally truncated Lig1 (residues 233 to 919; Fig. 1b) in the presence of ATP and Mg<sup>2+</sup>. The final step of DNA end-joining was blocked by having a 2', 3'-dideoxynucleoside at the 3' terminus of the nicked DNA. The crystal structure of the Lig1–AppDNA complex was determined by multiwavelength anomalous dispersion methods and the crystallographic model was refined against X-ray data extending to 3 Å resolution (Supplementary Table 1 and Supplementary Fig. S1).

The crystal structure shows that three domains of Lig1 completely encircle the AppDNA substrate (Fig. 2a, b), forming extensive interactions in the minor groove (Fig. 3a, b). The adenylation domain (AdD; residues 536 to 748) and the OB-fold domain (OBD; residues 749 to 919) comprise the catalytic core of Lig1 and resemble other members of a superfamily of covalent nucleotidyltransferases that includes ATP- and NAD<sup>+</sup>-dependent DNA ligases, RNA ligases and messenger RNA capping enzymes<sup>2</sup> (Fig. 1b). Although some minimal DNA ligases such as those of bacteriophage T7 (ref. 8) and *Chlorella* virus<sup>11</sup> catalyse efficient strand-joining using just the two-domain catalytic core, Lig1 has an additional N-terminal DNA-binding domain (DBD; residues 262 to 535) that is required for efficient ligation *in vitro* (Fig. 1c and Supplementary Table 2) and *in vivo*<sup>3</sup>. The DBD provides most of the DNA binding affinity (Supplementary Table 2) and it allows Lig1 to encircle its substrate through interactions with both the AdD and the OBD (Fig. 2a, b). All three domains of Lig1 cooperate to impose a sharp offset in the double helical axis that exposes the ends of the nicked DNA for interactions with active site residues and bound metals (Fig. 2c).

## A platform for DNA ligation

The DBD of Lig1 forms a broad, relatively flat surface that interacts with the minor groove of DNA (Fig. 3a, b). The twelve  $\alpha$ -helices of the DBD are arranged in a two-fold symmetric structure that interacts with the phosphodiester backbone on both sides of the nick (Fig. 3a). At each contact point, one DNA strand is engaged by a reverse turn located between two anti-parallel  $\alpha$ -helices and the other DNA strand is bound by an extended loop stretching between another pair of  $\alpha$ -helices (Fig. 3a). Most of these minor groove contacts are contributed by polar atoms of the protein main chain, whereas only a few basic side chains contact the DNA backbone. The DBD directly interacts with the AdD and the OBD of Lig1 (Fig. 2b) and stimulates the DNA end-joining activity of the catalytic core (Fig. 1c). The stimulation of ligation activity *in trans* may reflect interactions of the DBD that position the catalytic domains on DNA and/or the stabilization of the DNA substrate in a preferred orientation for ligation. The DBD is conserved in repair ligases III and IV (ref. 14), and many of the DBD residues contacting the other two domains of Lig1 are conserved (Supplementary Fig. S2). Truncations affecting the DBD of several eukaryotic DNA ligases severely impair enzymatic activity<sup>3,15–17</sup>. It is likely that all three mammalian DNA ligases have a ring-shaped architecture and recognize their DNA substrates in a similar manner.

The NAD<sup>+</sup>-dependent bacterial DNA ligases contain a helix–hairpin–helix (HhH) domain<sup>10</sup> that is analogous to the DBD, but is located carboxy-terminal to the catalytic core (Supplementary Fig. S3). Like the DBD, the HhH domain contributes most of the DNA binding affinity<sup>18</sup> and it has a two-fold symmetrical structure with the HhH DNA-binding elements properly spaced for interactions with the minor groove of DNA (Supplementary Fig. S3). An additional domain (Ia) that is required for enzyme–AMP formation is located N-terminal to the AdD of bacterial NAD<sup>+</sup>-dependent ligases<sup>19,20</sup>. This flexible N-terminal segment<sup>20</sup> is structurally similar to a region (residues 489 to 535) of the DBD abutting the AdD. Domain Ia might interact with the HhH domain, enabling bacterial ligases to encircle their DNA substrates<sup>10</sup> in a manner analogous to the Lig1–DNA complex.

## Active site structure

The AdD of Lig1 has a mixed  $\alpha$ ,  $\beta$  fold (Figs 2a and 3b) that closely resembles the nucleotide binding domains of bacterial and viral

DNA ligases<sup>8–11</sup> and mRNA capping enzymes<sup>21</sup>, consistent with the conservation of active site sequences<sup>2,12,13</sup>. In the Lig1–DNA complex, the 5' AMP cofactor is held outside of the DNA duplex, deep within a pocket of the AdD (Fig. 2a, c). Residues from conserved motifs I to V (Fig. 1b)<sup>2</sup> line the nucleotide binding pocket and interact with the 5' AMP to position the DNA ends and the ribose moiety of AMP (Fig. 2c). The conserved residues Glu 621 (motif III) and Arg 573 (motif I) are within hydrogen-bonding distance of the

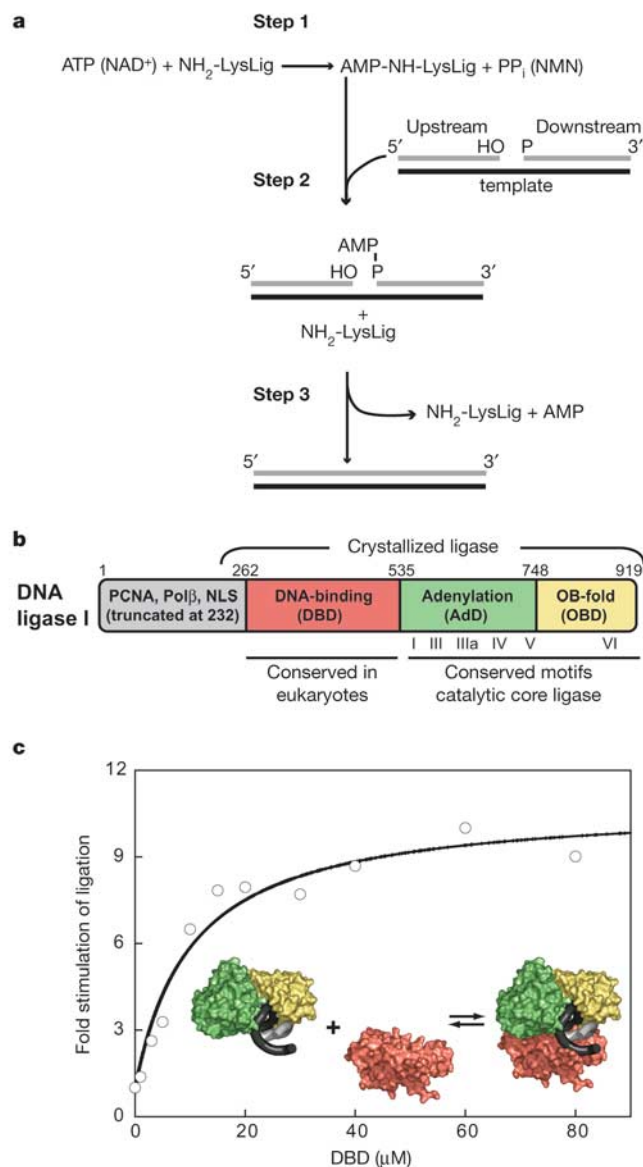
2'-OH and 3'-OH of AMP ribose, respectively, and these residues probably contribute to specificity for ribo-ATP (Fig. 2c and Supplementary Table 3). Lys 568 (motif I) forms a covalent enzyme–AMP adduct during step 1 of ligation (Fig. 1a) and its side chain lies near the 5'–5' pyrophosphate linkage of the AppDNA in the post-step 2 reaction intermediate that was crystallized (Fig. 2c). Lys 568 could promote DNA end-joining during step 3 (ref. 5) by correctly positioning the 5' phosphate and/or neutralizing the charge of the 5' AMP leaving group.

A hydrogen-bonding interaction between Glu 566 and the N6 of adenine (Fig. 2c) provides specificity for the ATP cofactor (Supplementary Table 3), and the side chain of Trp 742 (Fig. 2c) would additionally exclude the 2-amino group of GTP. Glu 566 is changed to Lys in one of the mutant *LIG1* alleles of a patient exhibiting severe immune deficiency<sup>22</sup>. Cells derived from this patient show delayed processing of Okazaki fragments and hypersensitivity to a variety of DNA damaging agents. The Glu566Lys mutation is predicted to strongly interfere with adenine binding to the AdD, providing a molecular explanation for the observed lack of enzymatic activity<sup>3</sup>. The other mutant *LIG1* allele in which Trp replaces Arg 771 has low activity and is defective in interacting with the DNA substrate<sup>22,23</sup>. In the crystal structure, Arg 771 is located on a loop (L12) in the OBD that binds to the template strand of DNA (Fig. 4b). Notably, a mouse model expressing this mutant version of Lig1 exhibits an increased predisposition to cancer<sup>24</sup>.

All DNA ligases are dependent upon divalent metal ions for catalysis, and the Lig1–DNA complex tentatively identifies two metal binding sites in the active site, supported by data obtained from X-ray experiments using crystals soaked in divalent metals and biochemical studies with ATP analogues. The weak electron density corresponding to the two metal binding sites (2.9  $\sigma$  and 3.2  $\sigma$  peaks in  $F_o - F_c$  difference maps) is suggestive of low binding occupancy (Fig. 2c). The 2', 3'-dideoxynucleoside in the crystal structure is not well aligned for nucleophilic attack of the 5' P of AppDNA (Fig. 2c). The absence of a 3'-OH may compromise metal binding and thereby interfere with the correct positioning of the DNA ends. One metal site is in close proximity to the AppDNA pyrophosphate linkage and is coordinated by Glu 720 (motif IV; Fig. 2c). This site is analogous to the metal binding site identified in crystal structures of other nucleotidyltransferases<sup>11,21</sup>. A metal at this position could stabilize the development of negative charge on the 5' phosphate of the AMP leaving group in the transition state for step 3 of the reaction (Fig. 1a). The other potential metal binding site identified in the Lig1–DNA complex is adjacent to the 2'-OH of AMP, and it is coordinated by the side chain of Glu 621 (motif III). This site is in close proximity to the (missing) 3'-OH of the nicked DNA (Fig. 2c), where it could participate in catalysis by positioning and activating the nucleophilic 3'-OH for attack on the 5' P. In support of this model, 2'-deoxy-2'-amino-ATP, in which the 2'-OH of AMP has been substituted with NH<sub>2</sub>, supports very low levels of ligation activity in comparison to ATP. Ligation activity with this analogue is partially rescued (stimulated more than tenfold) by the substitution of Mn<sup>2+</sup> for Mg<sup>2+</sup> (data not shown). Although ligation activity with 3'-deoxy-3'-amino ATP is also severely compromised in comparison to ATP, Mn<sup>2+</sup> does not affect activity with this analogue (not shown). The deleterious effect of the 3'-amino substitution could reflect a disruption of hydrogen bonding with Arg 573 (Fig. 2c). The coordinating residues (Glu 720 and Glu 621) at both metal binding sites are highly conserved and essential for efficient DNA ligation<sup>7</sup>, further supporting metals bound at these two locations (Fig. 2c). Additional structural and biochemical data will elucidate the roles of active site metal ions in the individual steps of the ligation reaction (Fig. 1a).

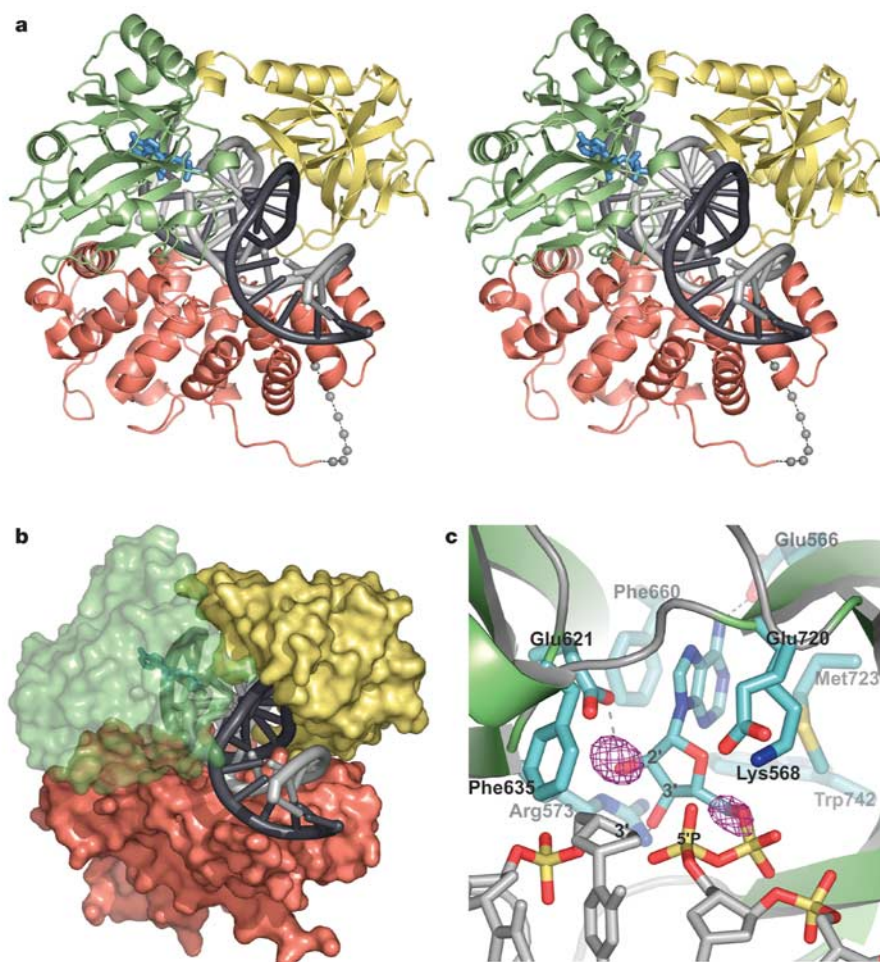
### DNA unwinding exposes the nick

The OBD binds in the minor groove adjacent to the ends of the nicked DNA (Figs 2a and 3b), and alters the curvature of the DNA



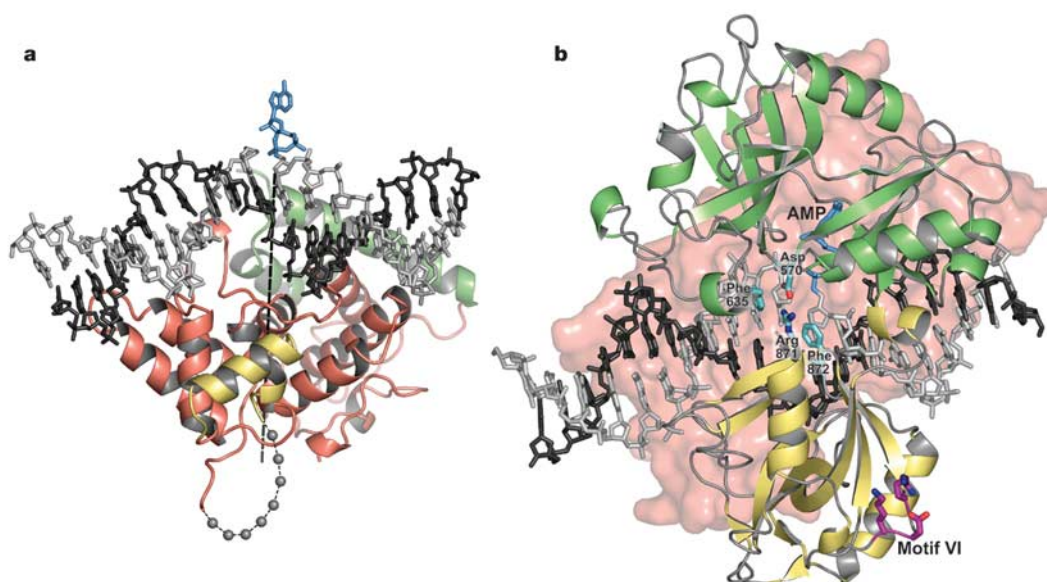
**Figure 1** Function and organization of Lig1. **a**, Enzymatic DNA ligation. Step 1, enzyme–AMP is formed by the attack of Lys on the  $\alpha$ -phosphate of ATP (or NAD<sup>+</sup>), releasing inorganic pyrophosphate (PP<sub>i</sub>) or nicotinamide mononucleotide (NMN). Step 2, the 5'-phosphate (5' P) of the nicked DNA strand (downstream) attacks the Lys–AMP intermediate to form an AppDNA intermediate (pyrophosphate linkage, 5' P to the 5' phosphate of AMP). Step 3, the 3'-OH terminated end of the nicked strand (upstream) attacks the 5' P of AppDNA, covalently joining the DNA strands and liberating AMP. **b**, Lig1 domain organization. The AdD (green) and the OBD (yellow) comprise the minimal catalytic core, harbouring motifs I to VI of the covalent nucleotidyltransferase superfamily. The DBD (red) is N-terminal to the catalytic core and is unique to eukaryotic ligases (Supplementary Fig. S2). An N-terminal region (grey) contains a nuclear localization signal (NLS)<sup>49</sup> and mediates protein interactions with polymerase  $\beta$  (Pol $\beta$ )<sup>50</sup> and PCNA<sup>38</sup>. **c**, Ligase activity of the catalytic core fragment is stimulated when the DBD is added *in trans* (Supplementary Table 2).





**Figure 2** Lig1 intimately engages its DNA substrate. **a**, Stereo view of the Lig1–DNA complex. Three domains of Lig1 (coloured as in Fig. 1b) fully encompass the AppDNA reaction intermediate. The DNA strands are coloured as in Fig. 1a, and the AppDNA linkage is drawn in blue. A poorly ordered surface loop (residues 385 to 392) was not modelled (grey spheres). **b**, Molecular surface of Lig1. The AdD is semi-transparent to

highlight the AMP cofactor held within the AdD active site. **c**, The AMP cofactor anchors the 5' P of the downstream DNA strand for interactions with catalytic residues. Two peaks of electron density from an  $F_o - F_c$  difference map (purple) mark the locations of two potential metal-binding sites.



**Figure 3** Lig1 engages the minor groove of DNA. **a**, The DBD binds the minor groove both upstream and downstream of the nick. An approximate two-fold axis of symmetry (dashed line) within the DBD is mirrored by the symmetry of the bound DNA. Segments of the DBD contact the OBD and the AdD (yellow and green regions, respectively). **b**, The OBD (yellow) and the AdD (green) interact to form a DNA-binding surface that places conserved

residues (Phe 635 and Phe 872) in the minor groove. A salt bridge (Asp 570 to Arg 871) stabilizes the AdD–OBD interface. The flat surface of the DBD (red) supports the DNA substrate and directly contacts the AdD and OBD. Motif VI residues (purple) face away from the AdD active site.



backbone, enforcing an underwound conformation that widens the major and minor grooves over a span of six base pairs (Fig. 4b). The DNA helical axes are offset by more than 5 Å on either side of the binding interface. DNA immediately upstream of the nick (as defined in Fig. 1a) adopts an A-form helix with an expanded minor groove, whereas the downstream DNA is in the B-form. A similar A-form to B-form transition is seen in DNA bound by the transcription factor Sac7d, which binds in the minor groove using a structurally similar  $\beta$ -barrel fold<sup>25</sup>. The OBD of Lig1 makes an extensive number of interactions with the DNA backbone that could stabilize the distorted conformation of the DNA substrate. Residues located on the surface of the OBD  $\beta$ -barrel in two interstrand connecting loops (L12 and L45) engage the template strand, and the N-terminal end of  $\alpha$ -helix S packs against the downstream side of the nicked DNA strand (Fig. 4b). These interactions with the template and substrate DNA strands specify a B-form width for the minor groove downstream of the nick (Figs 3b and 4b). During step 2, a similar asymmetric interaction with the DNA is indicated by the footprints of the bacteriophage T7 and *Chlorella* virus DNA ligases<sup>26,27</sup>. The A-form conformation of DNA upstream of the nick is stabilized by hydrophobic residues Phe 635 (Add) and Phe 872 (OBD) that broaden the minor groove and position the ends of the nicked DNA (Fig. 3b).

### Selection/recognition of ligation substrates

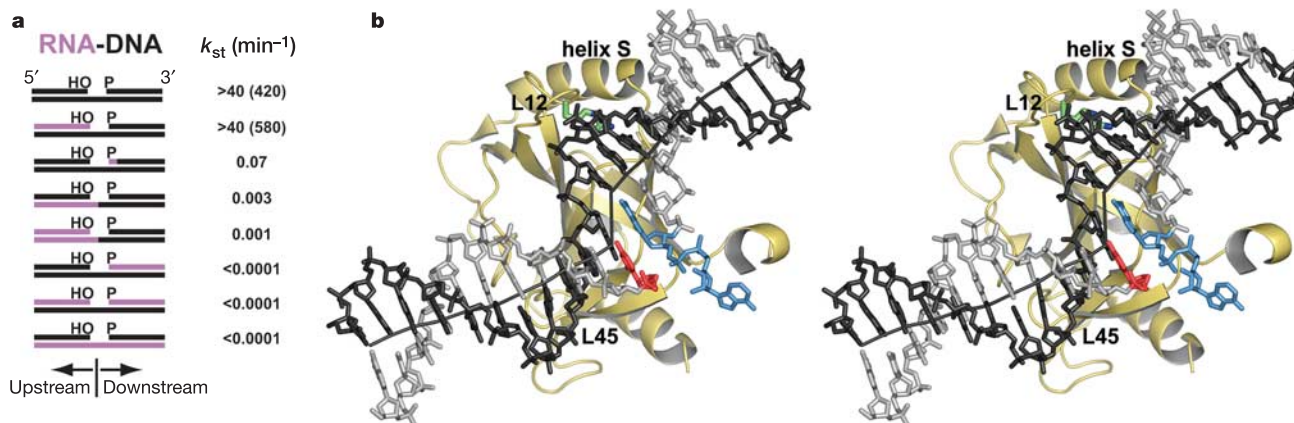
Several aspects of the fidelity of DNA ligation can be explained by the distorted conformation of DNA bound to Lig1. Most DNA ligases discriminate against substrates containing RNA strands<sup>28–30</sup> or mismatched base pairs at positions near the ends of the nicked DNA<sup>4,15,30–32</sup>. For Lig1, the discrimination against RNA-containing substrates prevents the untimely ligation of nascent Okazaki fragments before the 5' RNA primer is removed. RNA substituted on either strand downstream of the nick severely inhibits ligation activity *in vitro*, whereas an RNA strand is well tolerated upstream of the nick (Fig. 4a). In contrast to *Chlorella* virus DNA ligase<sup>30</sup>, Lig1 exhibits considerable discrimination against a single ribonucleotide at the 5' end of the nick (Fig. 4a)<sup>29</sup>. Much greater discrimination is observed against a substrate in which the 5' phosphorylated strand is completely RNA (Fig. 4a). There are no protein groups near the 2' H (OH) of the 5' nucleotide, suggesting that discrimination against downstream RNA relates to the shape or flexibility of an RNA:DNA heteroduplex<sup>29–31</sup>. The crystal structure shows that the A-form helical conformation of an RNA:DNA heteroduplex is incompatible

with the intimate interaction between the OBD and the B-form DNA downstream of the nick (Figs 3b and 4b). An RNA strand upstream of the nick is compatible with the A-form helix seen in the Lig1–AppDNA complex. In fact, a substrate with RNA in the upstream position of the nicked strand is ligated as efficiently as an all-DNA substrate (Fig. 4a).

The fidelity of ligation is also manifested by an intolerance towards DNA substrates with mispaired bases at the 3'-OH end of the nick—mispairs are generally better tolerated at the 5'-phosphorylated end of the nick<sup>4,15,30–32</sup>. The strong selection against mispaired bases on the 3' side of the nick may enhance the fidelity of excision repair of DNA damage by preventing the ligation of mismatches that arise during low fidelity repair synthesis<sup>32</sup>. The 5'-phosphorylated end of the DNA substrate is positioned in the active site of Lig1 by extensive interactions with the 5' AMP (Fig. 2c), diminishing the energetic importance of base pairing interactions at the 5' end of the nick. In contrast, the 3'-OH end makes few interactions with the enzyme and is correspondingly more dependent upon base pairing interactions with the template strand for proper alignment, explaining why 3' mispairs are more detrimental to ligation activity. The segment of the Add that positions Phe 635 in the minor groove upstream of the nick (Fig. 3b) might further ensure standard base pairing on the 3' end of the nick by recognizing proper hydrogen-bonding interactions in the minor groove<sup>33</sup>. The manner in which Lig1 fixes one end of a DNA break and manages the other end through base pairing interactions is reminiscent of the type I topoisomerases<sup>34</sup>. These enzymes encircle a nicked DNA substrate, holding one end tightly and permitting the other end to rotate. This mode of DNA binding can explain the topoisomerase-like relaxation of DNA by ligases<sup>35</sup>.

### Conformational switching during ligation

The orientation of the OBD in the Lig1–DNA complex highlights a large conformational change that must occur during ligation<sup>12</sup>. The motif VI residues that assist in enzyme–AMP formation (step 1)<sup>5,6</sup> and residues 871 and 872 that are involved in DNA–adenylate formation (step 2)<sup>6</sup> are located on opposite faces of the OBD  $\beta$ -barrel (Figs 3b and 5b). In the Lig1–DNA complex, residues 871 and 872 face the minor groove and the adjacent surface of the Add, in a location near the 5' P where they could assist with step 2 chemistry (Figs 3b and 5b)<sup>6</sup>. Arg 871 forms a salt bridge with Asp 570 (motif I) of the Add, orienting these domains as a continuous DNA binding surface (Fig. 3b). Phe 872 and Phe 635 (Add) are wedged into the



**Figure 4** Ligation fidelity. **a**, Ligation activity is greatly diminished for RNA:DNA heteroduplexes containing RNA (magenta) on either strand downstream of the nick; RNA is better tolerated upstream of the nick. The rate constants ( $k_{st}$ ) for single turnover reactions are listed. The values in parentheses are estimated from the observed rate constant at 4 °C and the relative rate constants at 4 °C and 21 °C for other substrates. **b**,

Stereo view of the OBD (yellow) as it distorts the DNA duplex, resulting in a A- to B-form transition of DNA structure across the nick (red to blue nucleotides). The DNA helical axis (black line) shifts by more than 5 Å at the nick site. The template strand (black) spans the length of the  $\beta$ -barrel between L12 and L45. Helix S abuts the downstream side of the nicked DNA strand. Arg 771 (green) extends from L12 into the minor groove.

minor groove and press against the ribose sugars of nucleotides on the 5' and 3' ends of the nick, aligning the reactants in the active site for ligation (Fig. 3b). It is instructive that motif VI residues functioning in the formation of ligase-AMP<sup>5,6</sup> are located on the exposed surface of the OBD away from the adenylate binding pocket in the Lig1-DNA complex (Figs 3b and 5b). Crystal structures of DNA ligases determined in the absence of DNA<sup>8,10,11</sup> reveal different orientations of the OBD and Add, indicative of a flexible linkage between domains<sup>12,36</sup>. The DNA binding residues (871 and 872) of the OBD are well separated from the ATP binding residues (motif VI; Fig. 3b), requiring two distinct modes of interaction between the OBD and the Add (Fig. 5a, b). Motif VI residues face the nucleotide binding pocket during step 1, as shown by the crystal structure of an

mRNA capping enzyme<sup>21</sup>. The alternative orientation of the OBD crystallized in complex with AppDNA is compatible with steps 2 and 3 of ligation (Fig. 5b).

### Interacting protein rings

Lig1 assembles as an active enzyme-DNA complex through direct interactions between its constituent domains and their complementary interactions with DNA that stabilize the ligation substrate in a nonstandard conformation. Although mammalian ligases III and IV are homologous to Lig1 and are predicted to interact with DNA in a similar manner, each enzyme has distinct cellular roles requiring specific interactions with other proteins<sup>14</sup>. The interaction of Lig1 with the proliferating cell nuclear antigen (PCNA) sliding clamp via a PCNA-interacting peptide motif<sup>37</sup> (the PIP box; residues 2 to 9) is critical for the joining of Okazaki fragments *in vivo*<sup>38</sup>. The similar sizes and annular shapes of Lig1 and PCNA can explain why only one Lig1 molecule binds to a PCNA trimer that is topologically linked to DNA<sup>39</sup>. An extended interface between these two stacked, ring-shaped proteins would occlude the interdomain connector loop regions of PCNA that bind the PIP box, thereby excluding other proteins from the complex (Supplementary Fig. S4). In support of this notion, the interaction of DNA ligase I with PCNA inhibits PCNA-dependent DNA synthesis by DNA polymerase  $\delta$  (ref. 39), and a region of the *Sulfolobus solfataricus* DNA ligase that is required for its interaction with PCNA<sup>40</sup> is homologous to residues 287 to 312 of the Lig1 DBD. Thus the crystal structure of human DNA ligase I in complex with nicked DNA not only provides molecular insight into the biochemical mechanism of ligation, but it also provides a framework for understanding the molecular mechanisms by which PCNA coordinates the processing and joining of Okazaki fragments. □

### Methods

#### Protein and DNA substrate preparation

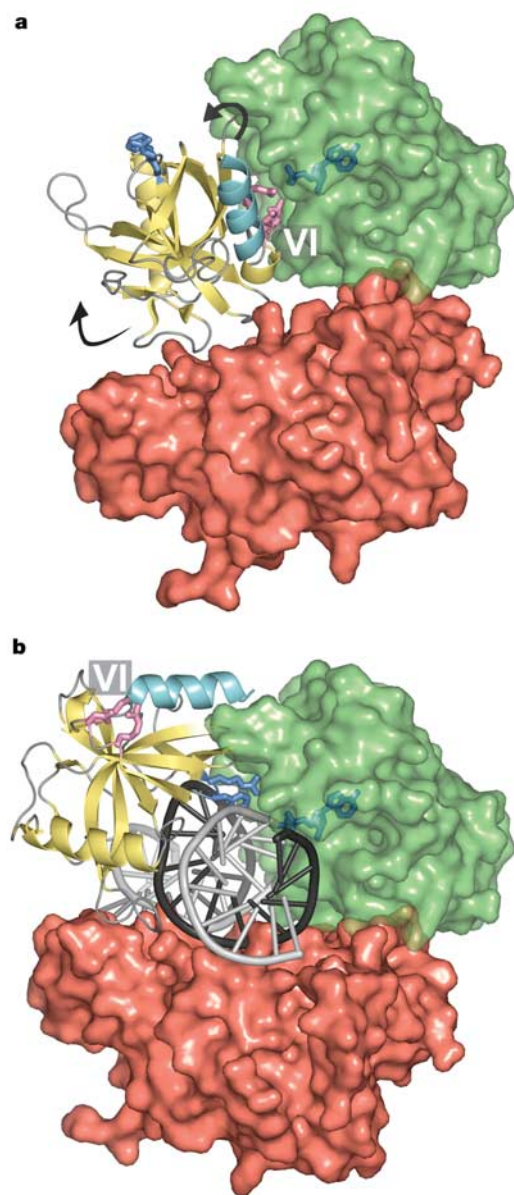
A full-length Lig1 expression vector<sup>41</sup> was altered to remove internal restriction sites *NdeI* and *HindIII*. The region coding for residues 233 to 919 was amplified by polymerase chain reaction (PCR) and subcloned into the *NdeI/HindIII* restriction sites of pET24b (Novagen). Selenomethionine (SeMet)-Lig1 was expressed in the *Escherichia coli* strain BL21(DE3)RP (Novagen) grown in defined media<sup>42</sup> and purified over phosphocellulose, HiTrap Q (Amersham), Cibacron Blue 3G (BioRad) and S200 (Pharmacia) gel filtration. Purified protein was concentrated in storage buffer (25 mM Tris pH 7.6, 150 mM NaCl, 0.1 mM EDTA and 5 mM DTT) to 25–45 mg ml<sup>-1</sup>. *NdeI/HindIII* fragments coding for residues 233 to 534 (DBD) or residues 533 to 919 (Add-OBD) were cloned into pET28a (Novagen). The N-terminally His<sub>6</sub>-tagged proteins were expressed in *E. coli* strain BL21(DE3)RP and purified using Ni<sup>2+</sup>-affinity, either cation (DBD) or anion (Add-OBD) exchange, and gel filtration chromatography. The polyhistidine tags were removed by cleavage with thrombin before gel filtration. Nicked DNA substrate was formed by annealing equimolar amounts of the three DNA strands [5'-(GTGCTGATGCGTddC)-3' (upstream; ddC is 2', 3' dideoxycytidine monophosphate), 5'-P-(GTGGGACTGATT CCG)-3' (downstream), and 5'-(CCGAATCAGTCCGACGACGCATCAGCAC)-3' (template)] in 5 mM MES pH 6.5 and 20 mM NaCl.

#### Lig1-DNA complex crystallization

A Lig1-DNA complex was formed by incubating 200  $\mu$ M ligase, 300  $\mu$ M DNA substrate, 1 mM ATP and 10 mM MgCl<sub>2</sub> at 30 °C for 10 min. The Lig1-DNA complex was mixed with an equal volume of well solution (5–7% polyethylene glycol (PEG) 4000, 100 mM sodium acetate pH 4.9, 5 mM DTT), and hexagonal crystals (*P*<sub>6</sub><sub>3</sub>; *a* = *b* = 161.9 Å and *c* = 88.5 Å) grew at 22 °C using sitting-drop vapour diffusion. Before flash-cooling in liquid nitrogen, crystals were transferred to a solution containing 4% PEG 4000, 30% ethylene glycol, 50 mM sodium acetate pH 4.9, 4 mM MgCl<sub>2</sub>, 25 mM NaCl and 1 mM DTT. Lig1-DNA crystals diffracted beyond 3 Å using synchrotron radiation, and there is one Lig1-DNA complex per asymmetric unit. A mercury (Hg) derivative was obtained by soaking the SeMet-Lig1 crystals for 12 h in cryo-solution supplemented with 0.2 mM methyl Hg acetate.

#### Structure determination

Multiwavelength anomalous dispersion (MAD) X-ray data were collected from two SeMet crystals and on one Hg derivative crystal (Supplementary Table 1) and processed using HKL2000 (ref. 43). Eight Hg sites were located by automated Patterson searches using SOLVE<sup>44</sup>. Six SeMet sites were located in anomalous difference Fourier maps constructed with Hg phases. Heavy-atom parameters were refined and experimental phases were calculated in SHARP<sup>45</sup>, treating SeMet data set 1 as native (Supplementary Table 1). Density modification was performed using the SOLOMON option in SHARP<sup>45</sup>. Two additional Hg derivative sites (ten in total) were located in residual density maps from



**Figure 5** Two active conformations of the OBD. **a**, Lig1 is modelled in a conformation competent for step 1 by superimposing the OBD (yellow) from Lig1 onto that of the mRNA capping enzyme<sup>21</sup> (PDB code 1CKM). The surface of the OBD bearing motif VI residues (VI; pink) faces the Add (green) active site. The DBD (red) must pivot ~20° to accommodate this conformation. **b**, OBD residues 871 and 872 (blue) face the active site during steps 2 and 3 (Lig1-DNA complex). In this conformation, motif VI residues are far from the active site. The relative positioning of the C-terminal  $\alpha$ -helix (cyan) highlights the rotation/movement of the OBD between these alternate conformations (compare **a** and **b**).



SHARP<sup>45</sup>. Experimentally phased maps had a well-defined solvent boundary and obvious protein and nucleic acid electron density. The electron density clearly showed the nick in the DNA duplex (Supplementary Fig. S1), establishing the register of the DNA nucleotide sequence. The Hg and SeMet sites, bulky amino-acid side chains, and the AdD and OBD of other DNA ligase structures<sup>8,11</sup> served as guides in registering the amino-acid sequence of Lig1.

The crystallographic model was constructed using O (ref. 46) and initially refined in CNS<sup>47</sup> to 3.0 Å resolution. In the later stages of model building, the structure was refined using REFMAC<sup>48</sup> with strict geometric and temperature factor restraints. TLS parameters were refined (REFMAC) with each of the three Lig1 domains and the DNA substrate treated as separate domains. The Lig1–DNA complex model contains 632 amino acids, 20 base pairs of duplex DNA and one AMP group with an  $R_{\text{cryst}}$  of 23.6% and an  $R_{\text{free}}$  of 26.8%. Eight residues were not observed in the electron density (residues 385 to 392), and the side chains of the following residues were not modelled beyond the C $_{\beta}$  atom: Arg 627, Glu 628, Arg 643, Lys 644, Glu 645, Val 646, Glu 709, Glu 712, Glu 807, Arg 859 and Lys 899. There are no Ramachandran violations, and Pro 477 is a *cis*-proline. The termini of the DNA substrate were poorly ordered in the crystal, and two base pairs on the upstream end and six base pairs on the downstream end were excluded from the model.

## DNA ligation assays

DNA oligonucleotides were synthesized using standard methods and RNA oligonucleotides were obtained from Dharmacon. Deprotection was performed according to the manufacturer's recommendations. The downstream (15mer) oligonucleotide was 5'-labelled with polynucleotide kinase. Reaction mixtures (10–20 µl) containing 50 mM Na MOPS pH 7.5, 1 mM DTT, 10 mM MgCl<sub>2</sub>, 50 µg ml<sup>-1</sup> BSA, 1–5 nM <sup>32</sup>P-labelled DNA or RNA substrate and enzyme were incubated at either 21 °C or 4 °C. Reactions were initiated by the addition of enzyme and halted by quenching a 2–4-µl aliquot in five volumes of 10 mM EDTA/formamide. Substrate (15mer), AppDNA intermediate (AppDNA–15mer), and ligation product (28mer) were separated on a 15% polyacrylamide gel containing 8 M urea and TBE (90 mM Tris-borate, 2.5 mM EDTA). The individual bands were quantified with a phosphorimaging system (Fuji BAS1000). The fractional extent of the ligation reaction was plotted as a function of time and the observed rate constant was obtained from an exponential nonlinear least squares fit to the data. For the slowest reactions ( $k < 0.003 \text{ min}^{-1}$ ) initial rates (<20% completion) were measured.

## Illustrations

Images were made using PyMol (<http://www.pymol.org>), Photoshop (Adobe Systems) and Illustrator (Adobe Systems).

Received 19 August; accepted 6 October 2004; doi:10.1038/nature03082.

- Lehman, I. R. DNA ligase: structure, mechanism, and function. *Science* **186**, 790–797 (1974).
- Shuman, S. & Schwer, B. RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol. Microbiol.* **17**, 405–410 (1995).
- Kodama, K., Barnes, D. E. & Lindahl, T. *In vitro* mutagenesis and functional expression in *Escherichia coli* of a cDNA encoding the catalytic domain of human DNA ligase I. *Nucleic Acids Res.* **19**, 6093–6099 (1991).
- Luo, J. & Barany, F. Identification of essential residues in *Thermus thermophilus* DNA ligase. *Nucleic Acids Res.* **24**, 3079–3085 (1996).
- Sriskanda, V. & Shuman, S. Mutational analysis of *Chlorella* virus DNA ligase: catalytic roles of domain I and motif VI. *Nucleic Acids Res.* **26**, 4618–4625 (1998).
- Mackey, Z. B. *et al.* DNA ligase III is recruited to DNA strand breaks by a zinc finger motif homologous to that of poly(ADP-ribose) polymerase. Identification of two functionally distinct DNA binding regions within DNA ligase III. *J. Biol. Chem.* **274**, 21679–21687 (1999).
- Sriskanda, V. & Shuman, S. Role of nucleotidyltransferase motifs I, III and IV in the catalysis of phosphodiester bond formation by *Chlorella* virus DNA ligase. *Nucleic Acids Res.* **30**, 903–911 (2002).
- Subramanya, H. S., Doherty, A. J., Ashford, S. R. & Wigley, D. B. Crystal structure of an ATP-dependent DNA ligase from bacteriophage T7. *Cell* **85**, 607–615 (1996).
- Singleton, M. R., Hakansson, K., Timson, D. J. & Wigley, D. B. Structure of the adenylation domain of an NAD<sup>+</sup>-dependent DNA ligase. *Struct. Fold. Des.* **7**, 35–42 (1999).
- Lee, J. Y. *et al.* Crystal structure of NAD(+)–dependent DNA ligase: modular architecture and functional implications. *EMBO J.* **19**, 1119–1129 (2000).
- Odell, M., Sriskanda, V., Shuman, S. & Nikolov, D. B. Crystal structure of eukaryotic DNA ligase-adenylate illuminates the mechanism of nick sensing and strand joining. *Mol. Cell* **6**, 1183–1193 (2000).
- Doherty, A. J. & Suh, S. W. Structural and mechanistic conservation in DNA ligases. *Nucleic Acids Res.* **28**, 4051–4058 (2000).
- Timson, D. J., Singleton, M. R. & Wigley, D. B. DNA ligases in the repair and replication of DNA. *Mutat. Res.* **460**, 301–318 (2000).
- Martin, I. V. & MacNeill, S. A. ATP-dependent DNA ligases. *Genome Biol.* **3**, Reviews 3005 (2002).
- Tomkinson, A. E., Tappe, N. J. & Friedberg, E. C. DNA ligase I from *Saccharomyces cerevisiae*: physical and biochemical characterization of the CDC9 gene product. *Biochemistry* **31**, 11762–11771 (1992).
- Sriskanda, V., Schwer, B., Ho, C. K. & Shuman, S. Mutational analysis of *Escherichia coli* DNA ligase identifies amino acids required for nick-ligation *in vitro* and for *in vivo* complementation of the growth of yeast cells deleted for CDC9 and LIG4. *Nucleic Acids Res.* **27**, 3953–3963 (1999).
- Grawunder, U., Zimmer, D. & Leiber, M. R. DNA ligase IV binds to XRCC4 via a motif located between rather than within its BRCT domains. *Curr. Biol.* **8**, 873–876 (1998).
- Jeon, H. J. *et al.* Mutational analyses of the thermostable NAD(+)–dependent DNA ligase from *Thermus filiformis*. *FEMS Microbiol. Lett.* **237**, 111–118 (2004).
- Sriskanda, V. & Shuman, S. Conserved residues in domain Ia are required for the reaction of *Escherichia coli* DNA ligase with NAD<sup>+</sup>. *J. Biol. Chem.* **277**, 9695–9700 (2002).
- Gajiwala, K. S. & Pinko, C. Structural rearrangement accompanying NAD(+) synthesis within a bacterial DNA ligase crystal. *Structure* **12**, 1449–1459 (2004).
- Hakansson, K., Doherty, A. J., Shuman, S. & Wigley, D. B. X-ray crystallography reveals a large

- conformational change during guanyl transfer by mRNA capping enzymes. *Cell* **89**, 545–553 (1997).
- Barnes, D. E., Tomkinson, A. E., Lehmann, A. R., Webster, A. D. & Lindahl, T. Mutations in the DNA ligase I gene of an individual with immunodeficiencies and cellular hypersensitivity to DNA-damaging agents. *Cell* **69**, 495–503 (1992).
- Prigent, C., Satoh, M. S., Daly, G., Barnes, D. E. & Lindahl, T. Aberrant DNA repair and DNA replication due to an inherited enzymatic defect in human DNA ligase I. *Mol. Cell. Biol.* **14**, 310–317 (1994).
- Harrison, C., Ketchen, A. M., Redhead, N. J., O'Sullivan, M. J. & Melton, D. W. Replication failure, genome instability, and increased cancer susceptibility in mice with a point mutation in the DNA ligase I gene. *Cancer Res.* **62**, 4065–4074 (2002).
- Robinson, H. *et al.* The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature* **392**, 202–205 (1998).
- Odell, M. & Shuman, S. Footprinting of *Chlorella* virus DNA ligase bound at a nick in duplex DNA. *J. Biol. Chem.* **274**, 14032–14039 (1999).
- Doherty, A. J. & Dafforn, T. R. Nick recognition by DNA ligases. *J. Mol. Biol.* **296**, 43–56 (2000).
- Sekiuchi, J. & Shuman, S. Ligation of RNA-containing duplexes by vaccinia DNA ligase. *Biochemistry* **36**, 9073–9079 (1997).
- Rumbaugh, J. A., Murante, R. S., Shi, S. & Bambara, R. A. Creation and removal of embedded ribonucleotides in chromosomal DNA during mammalian Okazaki fragment processing. *J. Biol. Chem.* **272**, 22591–22599 (1997).
- Sriskanda, V. & Shuman, S. Specificity and fidelity of strand joining by *Chlorella* virus DNA ligase. *Nucleic Acids Res.* **26**, 3536–3541 (1998).
- Shuman, S. Vaccinia virus DNA ligase: specificity, fidelity, and inhibition. *Biochemistry* **34**, 16138–16147 (1995).
- Bhagwat, A. S., Sanderson, R. J. & Lindahl, T. Delayed DNA joining at 3' mismatches by human DNA ligases. *Nucleic Acids Res.* **27**, 4028–4033 (1999).
- Liu, P., Burdzy, A. & Sowers, L. C. DNA ligases ensure fidelity by interrogating minor groove contacts. *Nucleic Acids Res.* **32**, 4503–4511 (2004).
- Corbett, K. D. & Berger, J. M. Structure, molecular mechanisms, and evolutionary relationships in DNA topoisomerases. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 95–118 (2004).
- Montecucco, A. & Ciarrocchi, G. AMP-dependent DNA relaxation catalyzed by DNA ligase occurs by a nicking-closing mechanism. *Nucleic Acids Res.* **16**, 7369–7381 (1988).
- Doherty, A. J. & Wigley, D. B. Functional domains of an ATP-dependent DNA ligase. *J. Mol. Biol.* **285**, 63–71 (1999).
- Warbrick, E. PCNA binding through a conserved motif. *Bioessays* **20**, 195–199 (1998).
- Levin, D. S., McKenna, A. E., Motycka, T. A., Matsumoto, Y. & Tomkinson, A. E. Interaction between PCNA and DNA ligase I is critical for joining of Okazaki fragments and long-patch base-excision repair. *Curr. Biol.* **10**, 919–922 (2000).
- Levin, D. S., Bai, W., Yao, N., O'Donnell, M. & Tomkinson, A. E. An interaction between DNA ligase I and proliferating cell nuclear antigen: implications for Okazaki fragment synthesis and joining. *Proc. Natl Acad. Sci. USA* **94**, 12863–12868 (1997).
- Dionne, I., Nookala, R. K., Jackson, S. P., Doherty, A. J. & Bell, S. D. A heterotrimeric PCNA in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol. Cell* **11**, 275–282 (2003).
- Teraoka, H. *et al.* Expression of active human DNA ligase I in *Escherichia coli* cells that harbor a full-length DNA ligase I cDNA construct. *J. Biol. Chem.* **268**, 24156–24162 (1993).
- Van Duyn, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229**, 105–124 (1993).
- Otinowski, Z. & Minor, W. in *Methods Enzymology* (eds Carter, C. W. & Sweet, R. M.) 307–326 (Academic, New York, 1997).
- Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 849–861 (1999).
- La Fortelle, E. D. & Bricogne, G. in *Methods Enzymology* (eds Sweet, R. M. & Carter, C. W.) 472–494 (Academic, New York, 1997).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
- Murshudov, G. N. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997).
- Montecucco, A. *et al.* The N-terminal domain of human DNA ligase I contains the nuclear localization signal and directs the enzyme to sites of DNA replication. *EMBO J.* **14**, 5379–5386 (1995).
- Dimitriadis, E. K. *et al.* Thermodynamics of human DNA ligase I trimerization and association with DNA polymerase  $\beta$ . *J. Biol. Chem.* **273**, 20540–20550 (1998).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** X-ray data were measured at beamlines X-12C and X-25 of the National Synchrotron Light Source (Upton, New York), and the SIBYLS beamline 12.3.1 of the Advanced Light Source (Berkeley, California), which are supported by the Offices of Biological and Environmental Research and of Basic Energy Sciences of the US Department of Energy, and the National Center for Research Resources of the National Institutes of Health. This work was supported by the Structural Cell Biology of DNA Repair Program Grant from the National Cancer Institute, and research grants from the National Institute of General Medical Sciences awarded to T.E. and A.E.T. J.M.P. and P.J.O. are supported by NRSA postdoctoral fellowships from the National Institutes of Health. T.E. is the Hsien Wu and Daisy Yen Wu Professor at Harvard Medical School.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to T.E. ([tome@hms.harvard.edu](mailto:tome@hms.harvard.edu)). Atomic coordinates and structure factor amplitudes have been deposited in the Protein Data Bank under accession code 1X9N.



# The building blocks of planets within the 'terrestrial' region of protoplanetary disks

R. van Boekel<sup>1,2</sup>, M. Min<sup>1</sup>, Ch. Leinert<sup>3</sup>, L.B.F.M. Waters<sup>1,4</sup>, A. Richichi<sup>2</sup>, O. Chesneau<sup>3</sup>, C. Dominik<sup>1</sup>, W. Jaffe<sup>5</sup>, A. Dutrey<sup>6</sup>, U. Graser<sup>3</sup>, Th. Henning<sup>3</sup>, J. de Jong<sup>5</sup>, R. Köhler<sup>3</sup>, A. de Koter<sup>1</sup>, B. Lopez<sup>7</sup>, F. Malbet<sup>6</sup>, S. Morel<sup>2</sup>, F. Paresce<sup>2</sup>, G. Perrin<sup>8</sup>, Th. Preibisch<sup>9</sup>, F. Przygodda<sup>3</sup>, M. Schöller<sup>2</sup> & M. Wittkowski<sup>2</sup>

<sup>1</sup>Astronomical Institute "Anton Pannekoek", University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

<sup>2</sup>European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching, Germany

<sup>3</sup>Max-Planck-Institut für Astronomie Heidelberg, Königstuhl 17, 69117 Heidelberg, Germany

<sup>4</sup>Instituut voor Sterrenkunde, K.U. Leuven, Celestijnenlaan 200B, 3001 Heverlee, Belgium

<sup>5</sup>Leiden Observatory, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands

<sup>6</sup>Observatoire de Bordeaux 2, rue de l'Observatoire F-33270 Floirac, France

<sup>7</sup>Observatoire de la Côte d'Azur, Département Fresnel UMR 6528, BP 4229, 06034 Nice Cedex 4, France

<sup>8</sup>Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique, Observatoire de Paris, section de Meudon, 5 place Jule Janssen, 92190 Meudon, France

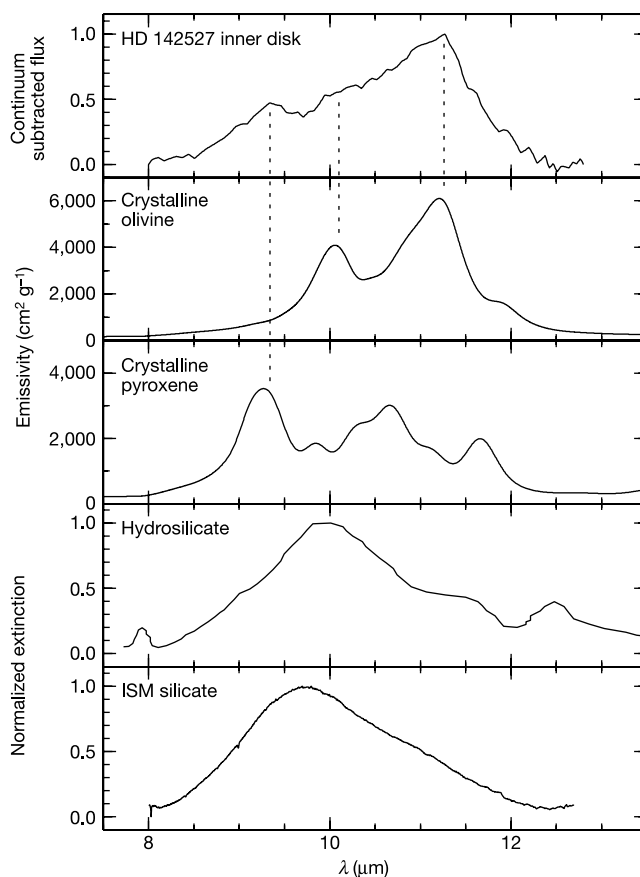
<sup>9</sup>Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany

Our Solar System was formed from a cloud of gas and dust. Most of the dust mass is contained in amorphous silicates<sup>1</sup>, yet crystalline silicates are abundant throughout the Solar System, reflecting the thermal and chemical alteration of solids during planet formation. (Even primitive bodies such as comets contain crystalline silicates<sup>2</sup>.) Little is known about the evolution of the dust that forms Earth-like planets. Here we report spatially resolved detections and compositional analyses of these building blocks in the innermost two astronomical units of three protoplanetary disks. We find the dust in these regions to be highly crystallized, more so than any other dust observed in young stars until now. In addition, the outer region of one star has equal amounts of pyroxene and olivine, whereas the inner regions are dominated by olivine. The spectral shape of the inner-disk spectra shows surprising similarity with Solar System comets. Radial-mixing models naturally explain this resemblance as well as the gradient in chemical composition. Our observations imply that silicates crystallize before any terrestrial planets are formed, consistent with the composition of meteorites in the Solar System.

Most young stars are surrounded by a disk of gas and dust which is a remnant of the star-formation process. This disk is formed owing to conservation of angular momentum in the collapsing proto-stellar cloud, and channels material from the cloud to the proto-star. When the material in the surrounding molecular cloud is exhausted, the disk dissipates within approximately 10<sup>7</sup> years (ref. 3). Planet formation is believed to result from the growth of submicrometre-sized interstellar dust particles<sup>4</sup>. Therefore, changes in size but also in the chemical nature of the dust grains in the nebular disk environment trace the first steps in planet formation. For instance, crystalline silicates are formed as a result of thermal annealing of amorphous grains, or by vaporization and subsequent gas-phase condensation in the innermost disk regions. These are referred to as primary processes. After inclusion of dust in larger parent bodies such as asteroids and planets, so-called secondary processing occurs, which includes oxidation, aqueous alteration and thermal metamorphism. Asteroids and comets contain pristine

interstellar dust as well as dust which has seen substantial processing<sup>5</sup>. The reconstruction of the formation history of our Solar System depends on a better understanding of the nature of primary and secondary processes, and when and where they occurred in the proto-solar nebula.

We observed three Herbig Ae stars with the Mid-Infrared Interferometric Instrument (MIDI)<sup>6</sup> installed at the Very Large Telescope Interferometer (VLTI). The light from two 8.2-m Unit Telescopes separated by 103 m on the ground was combined, providing a spatial resolution of about 20 milli-arcseconds. This corresponds to ~1–2 astronomical units (AU) at the distance of the observed stars; an improvement of more than a factor of ten in spatial resolution compared to the largest modern-day telescopes, in this wavelength regime. The MIDI instrument measures spectrally dispersed visibilities with  $\lambda/\Delta\lambda = 30$  in the 7.5–13.5- $\mu\text{m}$  atmospheric window. The intensity distribution of circumstellar disks is strongly centrally peaked<sup>7,8</sup>, so the correlated spectra measured by the interferometer are dominated by the inner 1–2 AU of the disks. We refer to these as the inner-disk spectra. In addition, spectra were obtained with a single 8.2-m telescope, in which the objects are spatially unresolved<sup>8</sup>. We refer to these spectra as the total-disk spectra. The difference between the total- and the inner-disk spectra arises mainly from a region between approximately 2 and 20 AU. We will refer to these spectra as the outer-disk spectra.



**Figure 1** The spectrum of the innermost disk regions of HD 142527 compared to spectra of typical dust species. From top to bottom we plot the observed inner-disk spectrum of HD 142527, the laboratory spectra of crystalline olivine and pyroxene<sup>29</sup>, a laboratory spectrum of an IDP consisting of hydrated silicates<sup>17</sup>, and the interstellar medium silicate spectrum<sup>1</sup>. The resolution of the laboratory data is reduced to that of the interferometric spectrum. The main resonances of crystalline pyroxene at 9.2  $\mu\text{m}$  and crystalline olivine at 11.3  $\mu\text{m}$  are clearly seen in the HD 142527 spectrum. We can exclude the possibility of a significant contribution of hydrated silicates to the spectrum in the inner-disk regions of HD 142527, which suggests that we see primary, rather than secondary dust.

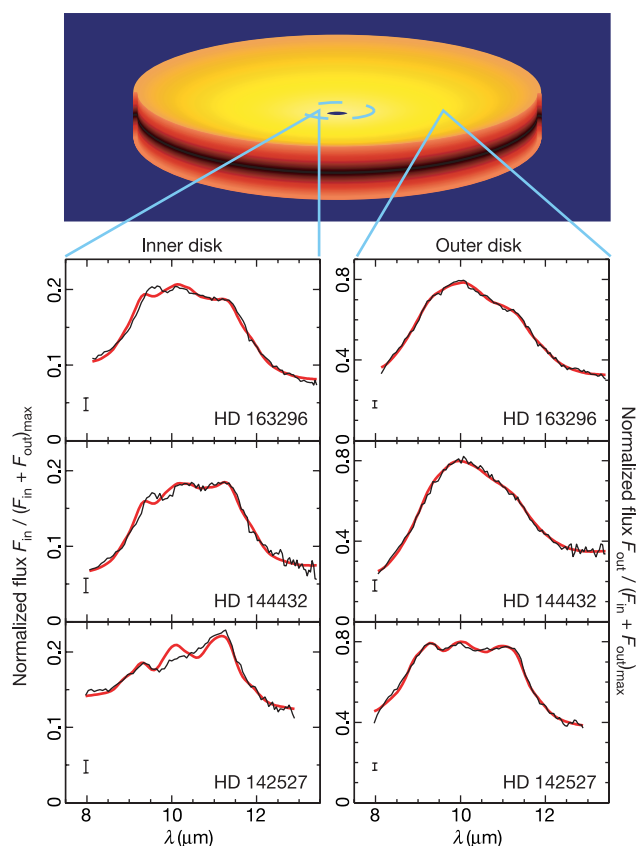
In Fig. 1 we show the inner-disk spectrum of HD 142527. The spectrum is clearly dominated by crystalline olivine and pyroxene. These minerals are also the most common in Solar System objects. No contribution from amorphous silicates is evident. As shown below, the spectrum is in fact consistent with 100 per cent crystalline material, making it the most crystalline dust ever observed in young stars<sup>9</sup>. Clearly, the mechanism responsible for crystallization must be highly efficient. Figure 2 shows the inner- and outer-disk spectra for all three stars. We have fitted the observations using laboratory measurements of materials that are the dominant species found in circumstellar disks<sup>10</sup> and interplanetary dust particles<sup>5</sup>. The derived mass fraction of crystalline silicates, the olivine over pyroxene ratio in the inner- and outer-disk spectra, and the fraction of material that resides in large grains, are listed in Table 1.

Both in the Solar System and in circumstellar disks crystalline silicates are found at large distances from the star. The origin of these silicates is a matter of debate. Although in the hot inner-disk

regions crystalline silicates can be produced by means of gas-phase condensation or thermal annealing, the typical grain temperatures in the outer-disk (2–20 AU) regions are far below the glass temperature of silicates of ~1,000 K. The crystals in these regions may have been transported outward through the disk<sup>11</sup> or in an outward-flowing wind<sup>12</sup>. An alternative source of crystalline silicates in the outer disk regions is *in situ* annealing, for example by shocks<sup>13</sup> or lightning<sup>14,15</sup>. A third way to produce crystalline silicates is the collisional destruction of large parent bodies in which secondary processing has taken place. We can use the mineralogy of the dust to derive information about the nature of the primary and/or secondary processes the small-grain population has undergone.

Models of disks accounting for the chemical equilibrium of a solid-gas mixture at high temperatures as well as the radial mixing of material<sup>11</sup> predict that the innermost disk region consists entirely of forsterite (the Mg-rich end member of the olivine family). At slightly lower temperatures, a reaction with gas-phase silicon efficiently converts this to enstatite (the Mg-rich end member of the pyroxene family). At a larger distance from the star, the predicted crystalline olivine over pyroxene ratio reaches about 0.5. Our observations qualitatively confirm these predictions. The inner-disk spectrum of HD 142527 imposes the strongest constraints because it shows the most-processed dust in our data. Assuming equal temperatures for all dust species, we find that the ratio of olivine to pyroxene in the inner-disk spectrum is 2.1, whereas it is 0.9 in the outer-disk spectrum. To our knowledge, this is the first direct measurement of a gradient in the chemical composition of the dust in proto-planetary disks. The measured gradient is smaller than is predicted by radial-mixing models. This could be due to non-equilibrium chemistry, or to the assumed stoichiometry in the amorphous silicates prior to annealing.

For all three stars the inner-disk regions have a substantially higher degree of crystallinity than the outer regions. However, these outer regions show a crystalline silicate abundance that clearly exceeds limits derived for the interstellar medium<sup>1</sup>. Therefore, these crystalline silicates must have been produced in the disk. It is not clear that the local processes that have been proposed to produce these crystals can account for the observed degree of crystallinity. Models for radial mixing of proto-planetary disks<sup>11</sup> can produce a crystalline fraction of several tens of per cent at distances of 5–10 AU (which is the relevant scale for our



**Figure 2** Infrared spectra of the inner (1–2 AU) and outer (2–20 AU) disk regions of three Herbig Ae stars. The outer-disk spectrum of each source has been constructed by subtracting the correlated spectrum from the total-disk spectrum (see ref. 8). The regions that dominate the inner- and outer-disk spectra are indicated in the schematic representation of a proto-planetary disk at the top of the figure (not to scale). The flux levels are scaled such that the sum of the inner- and outer-disk spectrum, that is, the total-disk spectrum, is normalized to unity. This allows the relative contributions of the inner and outer disk to the total spectrum to be estimated easily from this figure. The uncertainties in the spectra are indicated by the error bars in the lower left corner of each graph. The differences in shape between the inner- and outer-disk spectra are clearly visible in all three sources, indicating a difference in dust mineralogy. The broadening of the feature as seen in the inner-disk spectra indicates grain growth, whereas the resonance at 11.3  $\mu\text{m}$  indicates the presence of crystalline silicates (see also Fig. 1). Also shown are the best-fit model spectra for the inner- and outer-disk regions (red lines, see also Table 1). The model spectra reproduce the observed spectral shapes, although the fits to the inner disk spectra are less good than the fits to the outer-disk spectra.

**Table 1** Dust properties in the inner and outer disk

|           | Crystallinity (%)                |                                  | Fraction of large grains (%)     |                                  | Crystalline olivine to pyroxene ratio |                                     |
|-----------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|---------------------------------------|-------------------------------------|
|           | Inner disk                       | Outer disk                       | Inner disk                       | Outer disk                       | Inner disk                            | Outer disk                          |
| HD 163296 | 40 <sup>+20</sup> <sub>-10</sub> | 15 <sup>+10</sup> <sub>-10</sub> | 95 <sup>+5</sup> <sub>-10</sub>  | 65 <sup>+20</sup> <sub>-20</sub> | 2.3 <sup>+3.7</sup> <sub>-0.5</sub>   | –                                   |
| HD 144432 | 55 <sup>+30</sup> <sub>-20</sub> | 10 <sup>+10</sup> <sub>-5</sub>  | 90 <sup>+10</sup> <sub>-10</sub> | 35 <sup>+20</sup> <sub>-20</sub> | 2.0 <sup>+1.8</sup> <sub>-0.6</sub>   | –                                   |
| HD 142527 | 95 <sup>+5</sup> <sub>-15</sub>  | 40 <sup>+20</sup> <sub>-15</sub> | 65 <sup>+15</sup> <sub>-10</sub> | 80 <sup>+10</sup> <sub>-30</sub> | 2.1 <sup>+1.3</sup> <sub>-0.7</sub>   | 0.9 <sup>+0.2</sup> <sub>-0.1</sub> |

The fractional abundances of large and crystalline grains as well as the ratio of crystalline olivine to pyroxene in the inner (1–2 AU) and outer (2–20 AU) disk regions of three Herbig Ae stars. The dust components used in the model spectra are amorphous and crystalline olivine ( $\text{Mg}_{2-x}\text{Fe}_x\text{SiO}_4$ ), amorphous and crystalline pyroxene ( $\text{Mg}_{1-x}\text{Fe}_x\text{SiO}_3$ ), and amorphous silica ( $\text{SiO}_2$ ). For the crystalline components we used the magnesium-rich silicates ( $x = 1$ ) and for the amorphous components we used  $x = 0.5$ . The crystallinity is defined as the percentage of the total dust mass contributing to the 10- $\mu\text{m}$  feature, contained in crystalline olivine and pyroxene. Because the signature of crystalline silicates in the outer disk spectra of HD 163296 and HD 144432 is not very clear, the ratio of crystalline olivine over pyroxene in these spectra is not very well determined. The abundance of crystalline pyroxene in these spectra is particularly poorly constrained. All dust species are considered to have two different grain sizes, 0.1  $\mu\text{m}$  (small grains) and 1.5  $\mu\text{m}$  (large grains). Because we are looking at a coagulating environment where the various dust species are expected to be in thermal contact, we assume that all components have the same temperature. The opacities of the dust species are calculated from refractive indices obtained by laboratory measurements<sup>24–27</sup>, using the method of ref. 28. The abundances of the various dust species are determined using a linear least-squares fitting procedure assuming the dust emission comes from an optically thin part of the disk. The errors on the fitting parameters include both uncertainties in the data and in the modelling method, and are 1 $\sigma$ . As most of the error budget is due to systematic uncertainties that are the same in the inner and outer disk spectra, the relative difference between the two is much more significant than the error bars suggest. For details on the fitting procedure and the error analysis see the Supplementary Information.

observations), provided that a reservoir of crystalline material is present in the innermost disk regions. Our data prove that this reservoir exists.

Secondary processing of dust in large parent bodies produces hydrous silicates by means of aqueous alteration. Studies of meteorites have shown that the fine-grained matrix material of chondrites often consists of hydrous silicates<sup>16</sup>. However, we can exclude the possibility that the grains in our sources are predominantly hydrosilicates, because these show a fairly sharp peak near 10  $\mu\text{m}$ , as seen in interplanetary dust particles (IDPs) rich in hydrosilicates<sup>17</sup>. We conclude that aqueous alteration, followed by parent-body destruction, has not yet resulted in the production of a large abundance of small hydrosilicate grains in HD 142527. Therefore, parent-body processing can be excluded as the main source of crystalline silicates.

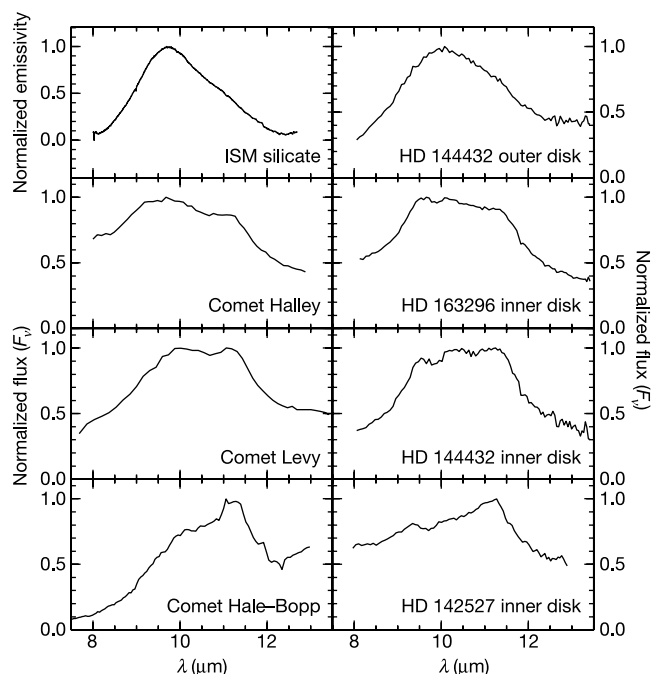
Figure 3 shows a remarkable similarity between the inner-disk spectra of our objects and those of Solar System comets, suggesting that the composition of the dust is also comparable. Therefore, the building blocks of comets in our Solar System have been processed in a similar way and to the same degree as in the inner disks of our programme stars. This is surprising, because comets formed in the icy regions of the solar nebula, further than 5 AU from the Sun. Cometary crystalline silicates are Mg-rich<sup>18</sup>. Chemical-equilibrium models indeed predict the formation of Mg-rich crystalline silicates at very high temperatures<sup>11</sup>. Measurements of the composition of

crystalline silicates that form in the outflows of red giants<sup>19</sup> support such chemical models. In addition, crystalline silicates found in fluffy, chondritic IDPs have a whisker or platelet morphology and internal crystallographic defects, indicative of gas-phase condensation<sup>5</sup>. This latter process is unlikely to occur in the outer regions of proto-planetary disks, but is expected to be important in the innermost regions. The most natural explanation for the presence of these materials in comets therefore seems to be transport by radial mixing from the rich reservoir of processed, Mg-rich crystalline silicate grains in the inner disk. Furthermore, radial-mixing models can account for the presence of cold crystalline silicates in the outer (> 20 AU) regions of some proto-planetary disks<sup>10,20</sup> as well as the occurrence of large (10–20  $\mu\text{m}$ ) FeS crystals in IDPs of cometary origin<sup>5</sup>.

It has been suggested that the degree of crystallinity of proto-planetary disks slowly increases with time after the accretion of matter on the star has stopped<sup>21</sup>. This evidence is based on the lack of crystalline silicates in infrared spectra of embedded young stars in the active phase, which are still accreting gas and dust from an interstellar cloud<sup>22</sup>. In contrast, passive proto-planetary disks around optically visible stars that are no longer accreting matter can show strong crystalline silicate emission<sup>20</sup>. The star with the highest abundance of crystalline silicates in our sample, HD 142527, is also the youngest one, with an age of approximately 1 million years (R.v.B. *et al.*, manuscript in preparation). Our observations thus imply that crystallization of almost the entire inner disk and a substantial part of the outer disk must have occurred very early in the evolution of the disk. Because dust processing—both radial mixing and shock processing—is more efficient during the active-disk phase, our observations provide strong evidence that crystallization occurred in the active-disk phase. The formation of the planets and asteroids in the inner Solar System is believed to have occurred on a much longer timescale<sup>23</sup>, during the passive disk phase. Therefore, our observations indicate that, as was the case in the early Solar System<sup>16</sup>, the silicate dust in the inner regions of proto-planetary disks is highly crystalline before planet formation occurs. □

Received 5 May; accepted 5 October 2004; doi:10.1038/nature03088.

- Kemper, F., Vriend, W. J. & Tielens, A. G. G. M. The absence of crystalline silicates in the diffuse interstellar medium. *Astrophys. J.* **609**, 826–837 (2004).
- Hanner, M. S., Lynch, D. K. & Russell, R. W. The 8–13 micron spectra of comets and the composition of silicate grains. *Astrophys. J.* **425**, 274–285 (1994).
- Haisch, K. E., Lada, E. A. & Lada, C. J. Disk frequencies and lifetimes in young clusters. *Astrophys. J. Lett.* **533**, L153–L156 (2001).
- Beckwith, S. V. W., Henning, T. & Nakagawa, Y. Dust properties and assembly of large particles in protoplanetary disks. In *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 533–558 (Univ. Arizona Press, 2000).
- Bradley, J. in *Lecture Notes in Physics* Vol. 609 *Astromineralogy* (ed. Henning, Th.) 217–235 (2003).
- Leinert, C. *et al.* Ten-micron instrument MIDI: getting ready for observations on the VLTI. In *Interferometry for Optical Astronomy II* (ed. Traub, W.A.). *Proc. SPIE* **4838**, 893–904 (2003).
- Eisner, J. A., Lane, B. F., Akeson, R. L., Hillenbrand, L. A. & Sargent, A. I. Near-infrared interferometric measurements of Herbig Ae/Be stars. *Astrophys. J.* **588**, 360–372 (2003).
- Leinert, C. *et al.* Mid-infrared sizes of circumstellar disks around Herbig Ae/Be stars measured with MIDI on the VLTI. *Astron. Astrophys.* **423**, 537–548 (2004).
- Molster, F. J. & Waters, L. B. F. M. in *Lecture Notes in Physics* Vol. 609 *Astromineralogy* (ed. Henning, Th.) 121–170 (2003).
- Bouwman, J. *et al.* Processing of silicate dust grains in Herbig Ae/Be systems. *Astron. Astrophys.* **375**, 950–962 (2001).
- Gail, H.-P. Radial mixing in protoplanetary accretion disks. IV. Metamorphosis of the silicate dust complex. *Astron. Astrophys.* **413**, 571–591 (2004).
- Nuth, J. A., Rietmeijer, F. J. M. & Hill, H. G. M. Condensation processes in astrophysical environments: The composition and structure of cometary grains. *Meteorit. Planet. Sci.* **37**, 1579–1590 (2002).
- Harker, D. E. & Desch, S. J. Annealing of silicate dust by nebular shocks at 10 AU. *Astrophys. J. Lett.* **565**, L109–L112 (2002).
- Pilipp, W., Hartquist, T. W., Morfill, G. E. & Levy, E. H. Chondrule formation by lightning in the Protosolar Nebula? *Astron. Astrophys.* **331**, 121–146 (1998).
- Desch, S. J. & Cuzzi, J. N. The generation of lightning in the solar nebula. *Icarus* **143**, 87–105 (2000).
- Brearely, A. & Jones, R. H. in *Planetary Materials* (ed. Papke, J.) Ch. 3 (The Mineralogical Society of America, 1998).
- Sandford, S. A. & Walker, R. M. Laboratory infrared transmission spectra of individual interplanetary dust particles from 2.5 to 25 microns. *Astrophys. J.* **291**, 838–851 (1985).



**Figure 3** A comparison between the spectral shapes of various astronomical objects (left column) with those of the inner- and outer-disk regions of our three Herbig Ae stars (right column). In the first row, we compare the silicate feature of the interstellar medium (ISM<sup>1</sup>) with the outer-disk spectrum of HD 144432. The silicate grains in the ISM are small (<0.1  $\mu\text{m}$ ) and amorphous. This results in a spectrum that has a typical triangular shape, peaking at 9.7  $\mu\text{m}$ . The outer-disk spectrum of HD 144432 is very similar, although a very weak shoulder at 11.3  $\mu\text{m}$  can be seen, indicating the presence of small amounts of processed material. In the second to the last rows, we compare the inner-disk spectra of our three Herbig stars with three Solar System comets<sup>2–18</sup>. The shapes of these spectra are very different from that of the ISM, indicating a higher degree of crystallinity, and on average larger dust grains. Grain growth makes the silicate emission band broader. Crystalline silicates have several bands in the 8–13- $\mu\text{m}$  spectral region, the strongest of which is at 11.3  $\mu\text{m}$ . Modest crystallinity causes the silicate feature to have rather sharp shoulders around 9.5 and 11.5  $\mu\text{m}$ , whereas in highly crystalline sources the crystalline silicates show prominent peaks (see Hale Bopp and HD 142527). The spectra are shown here in order of increasing crystallinity from top to bottom.



18. Crovisier, J. *et al.* The spectrum of Comet Hale-Bopp (C/1995 01) observed with the Infrared Space Observatory at 2.9 AU from the Sun. *Science* **275**, 1904–1907 (1997).
19. Molster, F. J., Waters, L. B. F. M., Tielens, A. G. G. M., Koike, C. & Chihara, H. Crystalline silicate dust around evolved stars. III. A correlations study of crystalline silicate features. *Astron. Astrophys.* **382**, 241–255 (2002).
20. Malfait, K. *et al.* The spectrum of the young star HD 100546 observed with the Infrared Space Observatory. *Astron. Astrophys.* **332**, L25–L28 (1998).
21. Grady, C. A. *et al.* Infalling planetesimals in pre-main stellar systems. In *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 613–638 (Univ. Arizona Press, 2000).
22. Acke, B. & van den Ancker, M. E. ISO spectroscopy of disks around Herbig Ae/Be stars. *Astron. Astrophys.* (in the press); preprint at (<http://xxx.lanl.gov/astro-ph/0406050>) (2004).
23. Wetherill, G. W. Formation of the earth. *Annu. Rev. Earth Planet. Sci.* **18**, 205–256 (1990).
24. Dorschner, J., Begemann, B., Henning, T., Jäger, C. & Mutschke, H. Steps toward interstellar silicate mineralogy. *Astron. Astrophys.* **300**, 503–520 (1995).
25. Servoin, J. L. & Piriou, B. Infrared reflectivity and Raman scattering of Mg<sub>2</sub>SiO<sub>4</sub> single crystal. *Phys. Status Solidi* **55**, 677–686 (1973).
26. Jäger, C. *et al.* Steps toward interstellar silicate mineralogy. IV. The crystalline revolution. *Astron. Astrophys.* **339**, 904–916 (1998).
27. Spitzer, W. G. & Kleinman, D. A. Infrared lattice bands of quartz. *Phys. Rev.* **121**, 1324–1335 (1961).
28. Min, M., Hovenier, J. W. & de Koter, A. Shape effects in scattering and absorption by randomly oriented particles small compared to the wavelength. *Astron. Astrophys.* **404**, 35–46 (2003).
29. Chihara, H., Koike, C. & Tsuchiyama, A. Low-temperature optical properties of silicate particles in the far-infrared region. *Publ. Astron. Soc. Jpn* **53**, 243–250 (2001).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** The data is based on observations obtained at the European Southern Observatory (ESO), Chile. We thank all those involved in building VLTI and MIDI. We thank V. Icke for providing the illustration shown in Fig. 2. C.P. Dullemond is acknowledged for many discussions.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to R.v.B. ([vboekel@science.uva.nl](mailto:vboekel@science.uva.nl)).

## Experimental demonstration of quantum memory for light

Brian Julsgaard<sup>1</sup>, Jacob Sherson<sup>1,2</sup>, J. Ignacio Cirac<sup>3</sup>, Jaromír Fiurášek<sup>4</sup> & Eugene S. Polzik<sup>1</sup>

<sup>1</sup>Niels Bohr Institute, Danish Quantum Optics Center – QUANTOP, Copenhagen University, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

<sup>2</sup>Department of Physics, Danish Quantum Optics Center – QUANTOP, University of Aarhus, 8000 Aarhus C, Denmark

<sup>3</sup>Max Planck Institute for Quantum Optics, Hans-Kopfermann-Str. 1, Garching, D-85748, Germany

<sup>4</sup>QUIC, Ecole Polytechnique, CP 165, Université Libre de Bruxelles, 1050 Brussels, Belgium, and Department of Optics, Palacky University, 17. listopadu 50, 77200 Olomouc, Czech Republic

The information carrier of today's communications, a weak pulse of light, is an intrinsically quantum object. As a consequence, complete information about the pulse cannot be perfectly recorded in a classical memory, even in principle. In the field of quantum information, this has led to the long-standing challenge of how to achieve a high-fidelity transfer of an independently prepared quantum state of light onto an atomic quantum state<sup>1–4</sup>. Here we propose and experimentally demonstrate a protocol for such a quantum memory based on atomic ensembles. Recording of an externally provided quantum state of light onto the atomic quantum memory is achieved with 70 per cent fidelity, significantly higher than the limit for classical recording. Quantum storage of light is achieved in three steps: first, interaction of the input pulse and an entangling field with spin-polarized caesium atoms; second, subsequent measurement of the transmitted light; and third, feedback onto the atoms using a radio-frequency magnetic pulse conditioned on the measure-

ment result. The density of recorded states is 33 per cent higher than the best classical recording of light onto atoms, with a quantum memory lifetime of up to 4 milliseconds.

Light is a natural carrier of information in both classical and quantum communications. In classical communications, bits are encoded in large average amplitudes of light pulses, which are detected, converted into electric signals and subsequently stored as charges or magnetization of memory cells. In quantum information processing, information is encoded in quantum states that cannot be accurately recorded by such classical means. Consider a state of light defined by its amplitude and phase, or equivalently by two quadrature phase operators,  $\hat{X}_L$  and  $\hat{P}_L$ , with the canonical commutation relation  $[\hat{X}_L, \hat{P}_L] = i$ . These variables play the same role in quantum mechanics as the classical quadratures  $X, P$  do in the decomposition of the electric field of light with the frequency  $\omega$  as  $E \propto X \cos \omega t + P \sin \omega t$ . Other quantum properties of light, such as the photon number  $\hat{n} = \frac{1}{2}(\hat{X}_L^2 + \hat{P}_L^2 - 1)$ , and so on, can be expressed in terms of  $\hat{X}_L$  and  $\hat{P}_L$ .

The best classical approach to recording a state of light onto atoms would involve homodyne measurements of both observables  $\hat{X}_L$  and  $\hat{P}_L$  by using, for example, a beam splitter. The non-commutativity of  $\hat{X}_L$  and  $\hat{P}_L$  leads to additional quantum noise being added during this procedure. The target atomic state has its intrinsic quantum noise (coming from the Heisenberg uncertainty relations). All this extra noise leads to a limited fidelity for the classical recording: for example, to a maximum fidelity of 50% for coherent states<sup>5–7</sup>. Thus the challenge of implementing a quantum memory can be formulated as a faithful storing of the simultaneously immeasurable values of  $\hat{X}_L$  and  $\hat{P}_L$ .

A number of quantum information protocols, such as eavesdropping in quantum cryptography, quantum repeaters<sup>8</sup>, and linear optics quantum computing<sup>9</sup>, would benefit from a memory meeting the following criteria: (1) the light pulse to be stored is sent by a third party in a state unknown to the memory party; (2) the state of light is converted into a quantum state of the memory with a fidelity higher than that of the classical recording. Several recent experiments<sup>10–13</sup> have demonstrated entanglement of light and atoms. However, none of these experiments demonstrated memory obeying the two above criteria. In ref. 14, where squeezed light was mapped onto atoms, the atomic state existed only while the light was on, so it was not a memory device. The electromagnetically induced transparency (EIT) approach has led to the demonstration of a classical memory for light<sup>15,16</sup>. A theoretical proposal for EIT-based quantum memory for light has been published in ref. 3. Other proposals for quantum memory for light with better-than-classical quality of recording have also been published recently<sup>1–4</sup>.

Quantum state transfer from one species to another is most simply presented if both systems are described by canonical quantum variables  $\hat{X}, \hat{P}$ . All canonical variables have the same commutation relations and the same quantum noise for a given state, thus providing a common frame for the analysis of the state transfer.

In the present work, the state of light is stored in the superposition of magnetic sublevels of the ground state of an atomic ensemble. As in ref. 12, we introduce the operator  $\hat{J}$  of the collective magnetic moment (orientation) of a ground state  $F$ . All atomic states utilized here are not too far in phase space from the coherent spin state (CSS), for which only one projection has a non-zero mean value, for example,  $\langle \hat{J}_x \rangle = J_x$ , whereas the other two projections have minimal quantum uncertainties,  $\langle \delta \hat{J}_y^2 \rangle = \langle \delta \hat{J}_z^2 \rangle = \frac{1}{2} J_x$ . For all such states, the commutator  $[\hat{J}_y, \hat{J}_z] = iJ_x$  can be reduced to the canonical commutator  $[\hat{X}_A, \hat{P}_A] = i$  with  $\hat{X}_A = \hat{J}_y/\sqrt{J_x}$ ,  $\hat{P}_A = \hat{J}_z/\sqrt{J_x}$ . Hence the  $y, z$ -components of the collective atomic angular momentum play the role of canonical variables. Although the memory protocol, in principle, can work with a single atomic ensemble, experimental technical noise is substantially reduced if two oppositely polarized ensembles placed in a bias magnetic field  $\mathbf{H}$  are used (see Methods

and Supplementary Methods for details). Combined canonical variables for two ensembles  $\hat{X}_A = (\hat{J}_{y1} - \hat{J}_{y2})/\sqrt{2J_x}$ ,  $\hat{P}_A = (\hat{J}_{z1} + \hat{J}_{z2})/\sqrt{2J_x}$  are then introduced, where  $\hat{J}_{x1} = -\hat{J}_{x2} = J_x = FN_{\text{atoms}}$ . In the presence of **H**, the memory couples to the  $\Omega$ -sidebands of light:  $\hat{X}_L = \frac{1}{\sqrt{T}} \int_0^T (\hat{a}^\dagger(t) + \hat{a}(t)) \cos(\Omega t) dt$ ,  $\hat{P}_L = \frac{i}{\sqrt{T}} \int_0^T (\hat{a}^\dagger(t) - \hat{a}(t)) \cos(\Omega t) dt$ , where  $\Omega$  is the Larmor frequency of spin precession.

Quantum storage of light is achieved in three steps: (1) an interaction of light with atoms; (2) a subsequent measurement of the transmitted light; and (3) feedback onto the atoms conditioned on the measurement result (Fig. 1). The off-resonant interaction of light with spin polarized atomic ensembles has been described elsewhere<sup>4,17–19</sup>, and is summarized in the Methods section. The interaction leads to the equations:

$$\begin{aligned}\hat{X}_L^{\text{out}} &= \hat{X}_L^{\text{in}} + k\hat{P}_A^{\text{in}}, & \hat{P}_L^{\text{out}} &= \hat{P}_L^{\text{in}} \\ \hat{X}_A^{\text{out}} &= \hat{X}_A^{\text{in}} + k\hat{P}_L^{\text{in}}, & \hat{P}_A^{\text{out}} &= \hat{P}_A^{\text{in}}\end{aligned}\quad (1)$$

These equations imply that light and atoms get entangled. The remarkable simplicity of equations (1) provides a direct link between an input light state, an atomic state, and an output light. Suppose the input light is in a vacuum (or in a coherent) state, and atoms are in a CSS with mean values  $\langle \hat{X}_L \rangle = \langle \hat{X}_A \rangle = \langle \hat{P}_L \rangle = \langle \hat{P}_A \rangle = 0$  and variances  $\delta X_L^2 = \delta X_A^2 = \delta P_L^2 = \delta P_A^2 = 1/2$ . The interaction parameter  $k$ , whose value is crucial for the storage protocol, is then readily found as  $k^2 = 2(\delta X_L^{\text{out}})^2 - 1$ .

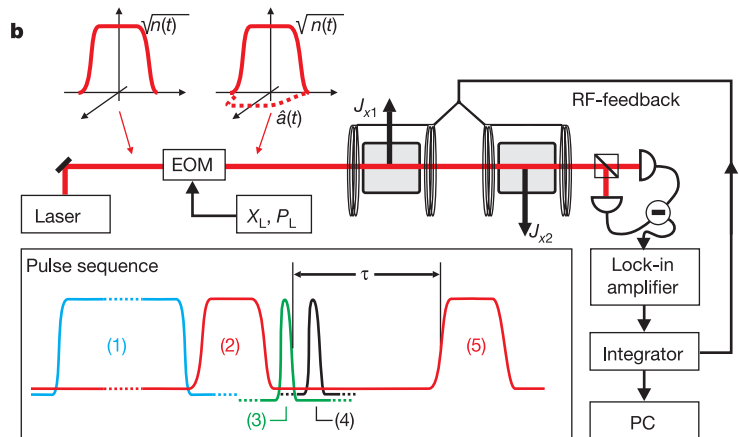
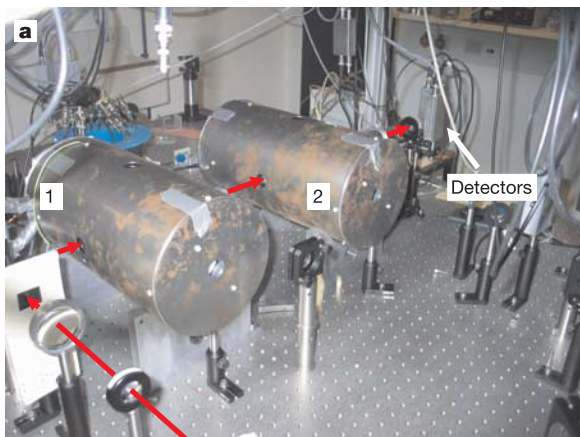
For a perfect fidelity of mapping, the initial atomic state must be an entangled spin state such as in ref. 12, with  $\delta X_A^2 \rightarrow 0$ . The pulse to be recorded, combined with the entangling pulse (see Methods section), is sent through, and its variable  $\hat{X}_L^{\text{out}}$  is measured. The measurement outcome,  $x = \hat{X}_L^{\text{in}} + k\hat{P}_A^{\text{in}}$ , is fed back into the atomic variable  $\hat{P}_A$  with a feedback gain  $g$ . The result is  $\hat{P}_A^{\text{mem}} = \hat{P}_A^{\text{in}} - gx = \hat{P}_A^{\text{in}}(1 - kg) - g\hat{X}_L^{\text{in}}$  (see Supplementary Notes for a justification of this equation). With  $g = k = 1$ , the mapping of  $\hat{X}_L^{\text{in}}$  onto  $-\hat{P}_A^{\text{mem}}$  is perfect.

The second operator of light is already mapped onto atoms via  $\hat{X}_A^{\text{mem}} = \hat{X}_A^{\text{in}} + \hat{P}_L^{\text{in}}$ , see equation (1). For the entangled initial state the mapping is perfect for this component too,  $\hat{P}_L^{\text{in}} \rightarrow \hat{X}_A^{\text{mem}}$ , leading to the fidelity of the light-to-atoms state transfer  $F \rightarrow 100\%$ . If the initial atomic state is a CSS, the mapping is not perfect owing to the noisy operator  $\hat{X}_A^{\text{in}}$ . However, fidelity  $F = 82\%$ , still markedly higher than the classical limit, can be achieved. Note that the above discussion holds for an arbitrary single mode input quantum state of light.

In our experiment, the atomic storage unit consists of two samples of caesium vapour placed in paraffin-coated glass cells placed inside magnetic shields (Fig. 1). **H** is applied along the  $x$ -direction with  $\Omega = 322$  kHz. Optical pumping along **H** initializes the atoms in the first/second sample in the  $F = 4$ ,  $m_F = \pm 4$  ground state with the orientation above 99%. Hence  $\hat{J}_{x1} = -\hat{J}_{x2} = J_x = 4N_{\text{atoms}} \approx 1.2 \times 10^{12}$ . We thoroughly check and regularly verify that the initial spin state is close to CSS (Supplementary Methods). The coupling parameter  $k$  is varied by adjusting the density of caesium vapour.

The input state  $\hat{a}(t)$  is encoded in a 1-ms  $y$ -polarized pulse. The state is chosen from the set  $\{\hat{a}_{\text{input}}\}$  of coherent states with the photon number in the range  $\{\langle n \rangle = 0, n_{\text{max}}\}$  and an arbitrary phase.  $\hat{a}(t)$  is generated as  $\Omega$  sidebands by an electro-optical modulator (EOM), and has the same spatial and temporal profile as the strong entangling field (more information can be found in the Methods section). Thus the EOM plays the third party, providing the field to be stored. The pulses are detuned by 700 MHz to the blue from the  $6S_{1/2}$ ,  $F = 4 \rightarrow 6P_{3/2}$ ,  $F = 5$  transition ( $\lambda = 852$  nm). The polarization measurement of the light is followed by the feedback onto atoms achieved by a 0.2 ms radio-frequency magnetic pulse conditioned on the measurement result.

The experimental verification of the quantum storage is then carried out. A read-out  $x$ -polarized pulse is sent through the samples with a delay of 0.7–10 ms after the feedback is applied. Atomic memory generates a  $y$ -polarized pulse, which is analysed as follows. As both  $\hat{X}_A^{\text{mem}}$  and  $\hat{P}_A^{\text{mem}}$  cannot be measured at the same time, we carry out two series of measurements for each input state. Each series consists of  $10^4$  quantum storage sequences. To verify the  $\hat{X}_L^{\text{in}} \rightarrow -\hat{P}_A^{\text{mem}}$  step of the storage, we measure the component  $\hat{X}_L^{\text{read-out}} = \hat{X}_L^{\text{read-in}} + k\hat{P}_A^{\text{mem}}$  of the read-out pulse ( $X_L$  is a Stokes parameter measured in units of shot noise, as discussed in the Methods section). An example of such a measurement carried out after 0.7 ms of storage is presented in Fig. 2a as a histogram of  $\frac{1}{k}\hat{X}_L^{\text{read-out}}$  (right histogram), with  $k$  measured as described in the Methods section and in Supplementary Methods. For this series  $\langle \hat{P}_L^{\text{in}} \rangle = -4$  and  $\langle \hat{X}_L^{\text{in}} \rangle = 0$ , corresponding to  $\langle \hat{n} \rangle = 8$  photons in the pulse. From this measurement, we find the mean  $\langle \hat{P}_A^{\text{mem}} \rangle = \frac{1}{k}(\hat{X}_L^{\text{read-out}})$  and the variance  $\sigma_p^2 = \langle (\delta \hat{P}_A^{\text{mem}})^2 \rangle = \frac{1}{k^2}((\delta \hat{X}_L^{\text{read-out}})^2 - \frac{1}{2})$  (see equations (1)) for the quantum state of the memory. We note that only the knowledge of  $k$  and the shot noise level of light is necessary for the determination of the mean values and variances of the atomic canonical variables from the experimental data.



**Figure 1** Experimental set-up. **a**, Atomic memory unit consisting of two caesium cells inside magnetic shields 1 and 2. The path of the recorded and read-out light pulses is shown with arrows. **b**, The simplified layout of the experiment. The input state of light with the desired displacements  $X_L$ ,  $P_L$  is generated with the electro-optical modulator (EOM). The inset shows the pulse sequence for the quantum memory recording and read-out.

Pulse (1) is the optical pumping (4 ms), pulse (2) is the input light pulse  $\hat{a}(t)$  overlapped with the strong entangling pulse in orthogonal polarization with amplitude  $\sqrt{n(t)}$ . Pulse (3) is the magnetic feedback pulse. Pulse (4) is the magnetic  $\pi/2$  pulse used for the read out of one of the atomic operators. Pulse (5) is the read-out optical pulse.

Next, we run another series of storage with the same input state for the verification of the step  $\hat{P}_L^{\text{in}} \rightarrow \hat{X}_A^{\text{mem}}$ . The  $\hat{X}_A^{\text{mem}}$  operator does not couple to the read-out pulse in our geometry; therefore, we first apply a  $\pi/2$ -pulse (Fig. 1) to atoms converting  $\hat{X}_A^{\text{mem}} \rightarrow \hat{P}_A^{\pi}$  and then measure  $\hat{P}_A^{\pi}$  with the verifying pulse. We then find  $\langle \hat{X}_A^{\text{mem}} \rangle$  and  $\sigma_x^2 = \langle (\delta \hat{X}_A^{\text{mem}})^2 \rangle$  of the memory state (left histogram).

The above sequence is repeated for different input states. From  $\langle \hat{P}_A^{\text{mem}} \rangle / \langle \hat{X}_L^{\text{in}} \rangle$  and  $\langle \hat{X}_A^{\text{mem}} \rangle / \langle \hat{P}_L^{\text{in}} \rangle$ , the mapping gains for the two quadratures are determined. For the experimental data presented in Figs 2 and 3a, these gains are 0.80 and 0.84 respectively, which is close to the optimal gain for the chosen input set of states. This step would complete the proof of the classical memory performance, because we have shown that the  $y$ -polarized pulse recovered from the memory has the same mean amplitude and mean phase as the input pulse (up to a chosen constant factor).

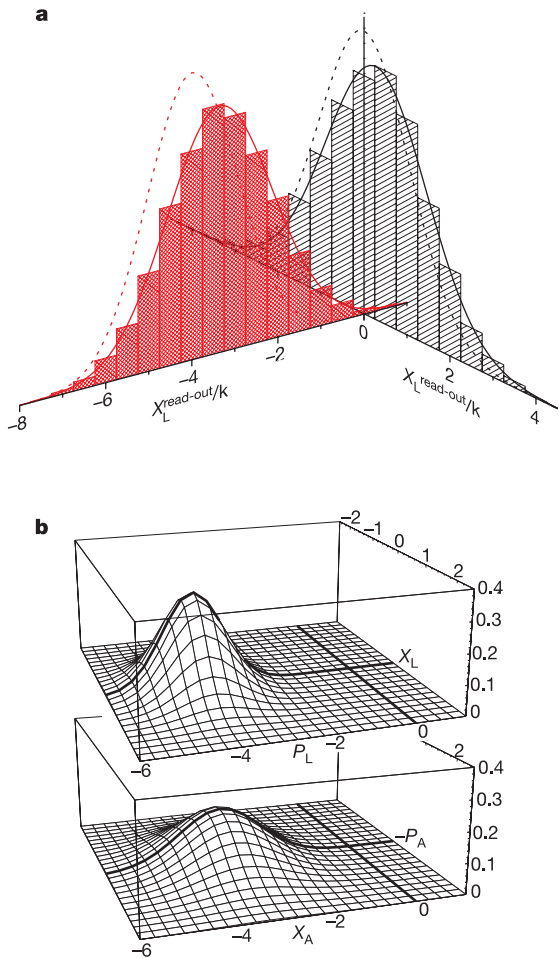
To prove a quantum memory performance, we need in addition to consider the quantum noise of the stored state. Towards this end, we plot the atomic variances  $\sigma_p^2, \sigma_x^2$  for the storage time 0.7 ms in Fig. 3a. The experimentally obtained variances of the stored state are on average 33% below the best possible variance of the classical recording. Hence a density of stored states 33% higher than that for the best classical recording can be obtained. Thus the goal of

quantum storage with less noise than for the classical recording is achieved.

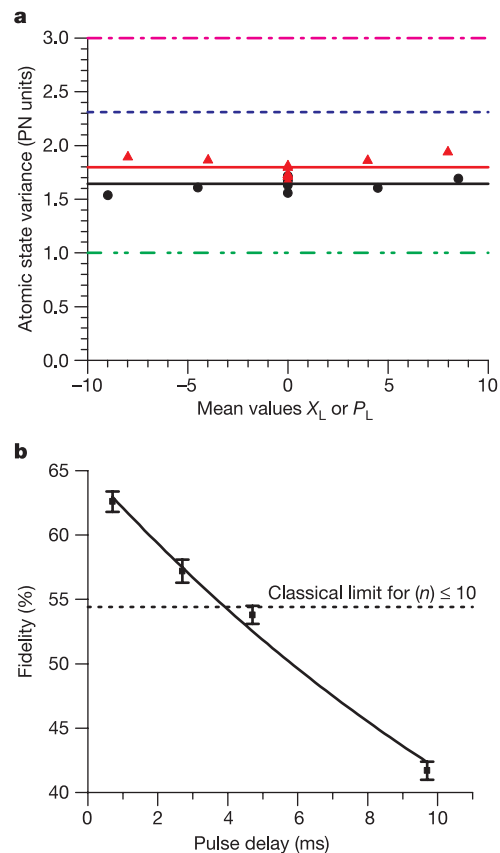
Next, the overlap between the input state of light and the state of the atomic memory is determined (Methods section). An example is shown in Fig. 2b. The fidelity  $F$  of the quantum recording is then calculated for a given set  $\{\hat{a}_{\text{input}}\}$ . For example,  $F = (66.7 \pm 1.7)\%$  for  $\{\hat{a}_{\text{input}}\} = \{n = 0 \rightarrow 8\}$  and  $F = (70.0 \pm 2.0)\%$  for  $\{\hat{a}_{\text{input}}\} = \{n = 0 \rightarrow 4\}$ , respectively, for the storage time of 0.7 ms. Note that the fidelity of the classical recording can exceed 50% for a limited set  $\{\hat{a}_{\text{input}}\}$ . The maximum classical fidelity for  $\{\hat{a}_{\text{input}}\} = \{n = 0 \rightarrow 8\}$  is 55.4%, and for  $\{\hat{a}_{\text{input}}\} = \{n = 0 \rightarrow 4\}$  it is 59.6%—still significantly lower than that for the quantum recording.

The main sources of imperfection of our quantum memory are decoherence of the atomic state and reflection off the cell walls. We have performed extensive studies of the atomic decoherence caused by the light-assisted collisional relaxation<sup>20</sup> to optimize the fidelity. Figure 3b presents the fidelity of the stored state as a function of the storage time. A simple model provides a good description for the observed fidelity reduction.

The single observable read-out described above can be useful in, for example, quantum cryptography eavesdropping, where the memory is read after the basis has been publicly announced by



**Figure 2** An example of the atomic memory performance. **a**, The input state of light in the coherent state with  $\langle \hat{X}_L \rangle = 0$ ,  $\langle \hat{P}_L \rangle = -4$ . The results of the read out of this state stored in the atomic memory are shown as histograms of experimental realizations. The left/right histogram shows the results for the  $\hat{X}_L/\hat{P}_A$  quadrature read out with/without the  $\pi/2$ -pulse. Dotted gaussians represent the distributions for the best possible quantum memory performance (fidelity 100%). **b**, The input coherent state of light (upper graph) and the reconstructed state stored in the atomic memory (lower graph) for the input state as in **a**. The reconstructed state is obtained from the results presented in **a** after subtracting the noise of the read-out pulse.



**Figure 3** Quantum noise of the stored state and the fidelity of quantum memory as a function of time. **a**, Experimental and theoretical (quantum and classical) stored state variances in atomic projection noise (PN) units. Triangles and filled circles are the experimental variances for the atomic memory operators, denoted  $2\sigma_x^2$  and  $2\sigma_p^2$ , respectively, in the text. Dash-dotted line, the fundamental boundary of three units of noise between quantum and classical mapping for an arbitrary coherent input state<sup>5,6</sup>. Dashed line, best classical variance for the experimental set of input states with photon numbers between 0 and 8. Double-dot-dashed line, unity variance corresponding to perfect mapping. **b**, Fidelity as a function of storage time for the set of states from 0 to 10 photons. Fidelity higher than the classical limit is maintained for up to 4 ms of storage. Error bars are standard deviations.



Alice and Bob. The present experiment also paves the way towards the proposed quantum cloning of light onto atomic memory<sup>21</sup>. However, other applications require complete state recovery via reverse mapping of the memory state onto light. Proposals for performing this task within our approach have been published<sup>14,19,22</sup>. Probably the most intuitively clear protocol for the memory read-out is just to run the storage protocol presented here with reversed roles of light and atoms. Indeed, the equations of interaction (1) are completely symmetric. The read-out, as the storage, would involve three steps: sending a read-out light pulse through atoms, measuring the spin projection  $\hat{X}_{\text{out}}^{\text{out}}$  with an auxiliary light pulse, and applying the feedback conditioned on this measurement to the read-out pulse.

In the present experiment, we have demonstrated the memory for a subset of linearly independent coherent states. Owing to the linearity of quantum mechanics, this demonstration signifies that our method provides faithful mapping for an arbitrary coherent state. As any arbitrary quantum state can be written as a superposition of coherent states, our approach should in principle work for an arbitrary quantum state, including entangled and single photon (qubit) states. □

## Methods

### Quantum coupling of light to two atomic ensembles in the presence of magnetic field

Here we discuss the physics behind the equations of interaction (1). The off-resonant atom/light interaction is described in terms of Stokes operators for the polarization state of light and the collective spin of atoms<sup>4,17,18</sup>. The Stokes operators are defined as one half of the photon number difference between orthogonal polarization modes:  $\hat{S}_1$ , between vertical  $x$ - and horizontal  $y$ -polarizations;  $\hat{S}_2$ , between the modes polarized at  $\pm 45^\circ$  to the vertical axis; and  $\hat{S}_3$ , between the left- and right-hand circular polarizations. In the experiment, a strong entangling  $x$ -polarized pulse with photon flux  $n(t)$  is mixed on a polarizing beamsplitter with the  $y$ -polarized quantum field  $\hat{a}(t)$  prior to interaction with atoms. Hence the Stokes operators of the total optical field are  $\hat{S}_1(t) = \hat{S}_1(t) = \frac{1}{2}n(t)$ ,  $\hat{S}_2 = \frac{1}{2}\sqrt{n(t)}(\hat{a}^+(t) + \hat{a}(t))$ ,  $\hat{S}_3 = \frac{i}{2}\sqrt{n(t)}(\hat{a}^+(t) - \hat{a}(t))$ . Note that  $\hat{S}_2(t)$  and  $\hat{S}_3(t)$  are proportional to the canonical variables for the quantum light mode,  $\hat{X} = \frac{1}{\sqrt{2}}(\hat{a}^+(t) + \hat{a}(t))$ ,  $\hat{P} = \frac{i}{\sqrt{2}}(\hat{a}^+(t) - \hat{a}(t))$ . Light is transmitted through the atomic samples placed in the bias magnetic field oriented along the  $x$ -axis. The magnetic field allows for encoding of the memory at the Larmor frequency  $\Omega$ , thus dramatically reducing technical noise present at low frequencies. However, in the presence of the Larmor precession, there is an undesired coupling of the single-cell variables  $\hat{J}_y$  and  $\hat{J}_z$  to each other. The introduction of the second cell with the opposite Larmor precession allows us to introduce new two-cell variables ( $\hat{J}_{y1} - \hat{J}_{y2}$ ), ( $\hat{J}_{z1} + \hat{J}_{z2}$ ) that do not couple to each other. As in ref. 12, where a similar trick was used, the Stokes parameters of light transmitted through the two cells along the  $z$  direction become:

$$\hat{S}_2^{\text{out}}(t) = \hat{S}_2^{\text{in}}(t) + a\hat{S}_1(\cos(\Omega t)[\hat{J}_{z1} + \hat{J}_{z2}] + \sin(\Omega t)[\hat{J}_{y1} + \hat{J}_{y2}]), \quad \hat{S}_3^{\text{out}}(t) = \hat{S}_3^{\text{in}}(t) \quad (2)$$

where  $\hat{J}_{z,y}$  are the projections in the frame rotating at  $\Omega$ , and  $a = \frac{\gamma\lambda^2}{8\pi\Delta A}$ , with  $\gamma$  and  $\lambda$  the natural linewidth and the wavelength of the transition, respectively,  $\Delta$  the detuning, and  $A$  the beam cross-section. At the same time, the transverse spin components of the two cells evolve as follows:

$$\begin{aligned} \frac{d}{dt}[\hat{J}_{z1} + \hat{J}_{z2}] &= \frac{d}{dt}[\hat{J}_{y1} + \hat{J}_{y2}] = 0, \\ \frac{d}{dt}[\hat{J}_{y1} - \hat{J}_{y2}] &= 2a\hat{J}_x\hat{S}_3^{\text{in}}\cos(\Omega t), \quad \frac{d}{dt}[\hat{J}_{z1} - \hat{J}_{z2}] = 2a\hat{J}_x\hat{S}_3^{\text{in}}\sin(\Omega t) \end{aligned} \quad (3)$$

As evident from equation (3), in the process of propagation the operator  $\hat{S}_3^{\text{in}}$  is recorded onto the operators  $\hat{J}_{y1} - \hat{J}_{y2}$  and  $\hat{J}_{z1} - \hat{J}_{z2}$  (the 'back action' of light on atoms via the dynamic Stark effect caused by light<sup>17,18</sup>), while the operators  $\hat{J}_{y1} + \hat{J}_{y2}$  and  $\hat{J}_{z1} + \hat{J}_{z2}$  are left unchanged. The latter are read out onto  $\hat{S}_2^{\text{out}}$  via the Faraday rotation, equation (2).

Canonical variables are defined for the quantum light mode as  $\hat{X}_L = \frac{1}{\sqrt{T}}\int_0^T \hat{a}^+(t) + \hat{a}(t)\cos(\Omega t)dt$ ,  $\hat{P}_L = \frac{i}{\sqrt{T}}\int_0^T (\hat{a}^+(t) - \hat{a}(t))\cos(\Omega t)dt$ ; that is, the relevant light mode involves the  $\Omega$ -sidebands.  $T$  is the pulse duration,  $\hat{a}(t)$  is normalized to the photon flux.  $\hat{X}_L$  and  $\hat{P}_L$  (that is,  $\hat{S}_2$  and  $\hat{S}_3$ ) are detected by a polarization state analyser and by lock-in detection of the  $\Omega$  component of the photocurrent. Note that the  $\cos(\Omega t)$  component of light couples to the ( $\hat{J}_{y1} - \hat{J}_{y2}$ ), ( $\hat{J}_{z1} + \hat{J}_{z2}$ ) components of atomic storage variables (equations (2), (3)). The equivalent choice of a  $\sin(\Omega t)$  modulation instead would mean the use of ( $\hat{J}_{y1} + \hat{J}_{y2}$ ), ( $\hat{J}_{z1} - \hat{J}_{z2}$ ) for the memory. The atomic canonical variables  $\hat{X}_A$ ,  $\hat{P}_A$  are defined in the main section. With the above equations and definitions we straightforwardly derive equations (1) under the assumption  $\Omega T \gg 1$ . Theoretically, the dimensionless coupling parameter in (1) is  $k^2 = \frac{1}{2}a^2\hat{J}_x \int n(t)dt$ .

### Experimental calibration of the canonical variances for light and atoms

Calculations of the fidelity, the gains, and the variances from the experimental data are based on the experimental calibration of  $\langle \delta X_L^2 \rangle = \langle \delta P_L^2 \rangle$  for the coherent (vacuum) state of light and of  $\langle \delta X_A^2 \rangle = \langle \delta P_A^2 \rangle$  for the CSS of atoms. The calibration for light is carried out along the established procedure of determining the shot noise level for measurements of

$\hat{S}_2$ ,  $\hat{S}_3$  with the quantum field in a vacuum state<sup>5,17</sup>. Variances and mean values for light are then measured in units of this shot noise level. The calibration for the atomic CSS variance is carried out with extreme care, and has shown excellent reproducibility (see Supplementary Methods). As stated in the main text, as soon as the vacuum (shot) noise level for light is established and the atoms are in a CSS, the parameter  $k^2$  (equations (1)), important for calculations of atomic variances and fidelity, is easily determined as  $k^2 = 2(\delta X_L^{\text{out}})^2 - 1$ . In the experiment, this is equivalent to  $k^2 = ((\delta S_2^{\text{out}})^2 - (\delta S_2^{\text{in}})^2) / (\delta S_2^{\text{in}})^2$ .

### Fidelity and the state overlap

To calculate the fidelity of the transfer of an input coherent state into an output gaussian state<sup>8</sup>, we first define an overlap function between an input state with mean values  $x_1$ ,  $p_1$  and the output state with the mean values and variances  $x_2$ ,  $p_2$ ,  $\sigma_x^2$ ,  $\sigma_p^2$ . Straightforward integration yields:

$$\begin{aligned} O\{x_1, x_2, p_1, p_2\} &= 2\exp(-(x_1 - x_2)^2 / (1 + 2\sigma_x^2)) \\ &\quad - (p_1 - p_2)^2 / (1 + 2\sigma_p^2)) / \sqrt{(1 + 2\sigma_x^2)(1 + 2\sigma_p^2)} \end{aligned}$$

The fidelity of the transfer for a set of coherent states with mean amplitudes between  $\alpha_1$  and  $\alpha_2$  can then be found as an average overlap:

$$F = \pi^{-1}(\alpha_2^2 - \alpha_1^2)^{-1} \int_0^{2\pi} d\phi \int_{\alpha_1}^{\alpha_2} O\{\alpha\} \alpha d\alpha$$

For classical recording from light onto atoms with gain  $g$ , the overlap between the input coherent state with the mean amplitude  $\alpha = \sqrt{x^2 + p^2}$  and the output state is given by  $O\{\alpha\} = (1 + g^2)^{-1} \exp(-\frac{1}{2}(1 - g^2)\alpha^2(1 + g^2)^{-1})$ . The classical fidelity is then given by:

$$\begin{aligned} F_{\text{class}} &= (n_2 - n_1)^{-1}(1 - g)^{-2} \exp(-(1 - g^2)n_1(1 + g^2)^{-1}) \\ &\quad - \exp(-(1 - g^2)n_2(1 + g^2)^{-1}) \end{aligned}$$

where we have introduced the mean photon number  $n = \frac{1}{2}\alpha^2$ .  $F_{\text{class}} \rightarrow 50\%$  for arbitrary coherent states when  $g \rightarrow 1$ . If a restricted class of coherent states is chosen as the input,  $F_{\text{class}} > 50\%$  can be obtained with a suitable choice of  $g$ . For a set of states analysed in the main text,  $\{d_{\text{input}}\} = \{n = 0 \rightarrow 8\}$ , the maximum classical fidelity of 55.4% is achieved with a gain of 0.809.

Received 11 June; accepted 28 September 2004; doi:10.1038/nature03064.

1. Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).
2. Kozhekin, A. E., Mølmer, K. & Polzik, E. S. Quantum memory for light. *Phys. Rev. A* **62**, 033809 (2000).
3. Fleischhauer, M. & Lukin, M. D. Quantum memory for photons: dark-state polaritons. *Phys. Rev. A* **65**, 022314 (2002).
4. Kuzmich, A. & Polzik, E. S. In *Quantum Information with Continuous Variables* (eds Braunstein, S. L. & Pati, A. K.) 231–265 (Kluwer, Dordrecht, 2003).
5. Furusawa, A. *et al.* Unconditional quantum teleportation. *Science* **282**, 706–709 (1998).
6. Van Loock, P., Braunstein, S. L. & Kimble, H. J. Broadband teleportation. *Phys. Rev. A* **62**, 022309 (2000).
7. Hammerer, K., Wolf, M. M., Polzik, E. S. & Cirac, J. I. Quantum benchmark for storage and transmission of coherent states. Preprint at (<http://xxx.lanl.gov/quant-ph/0409109>) (2004).
8. Briegel, H. J., Dur, W., Cirac, J. I. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
9. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–53 (2001).
10. van der Wal, C. H. *et al.* Atomic memory for correlated photon states. *Science* **301**, 196–200 (2003).
11. Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* **423**, 731–734 (2003).
12. Julsgaard, B., Kozhekin, A. & Polzik, E. S. Experimental long-lived entanglement of two macroscopic objects. *Nature* **413**, 400–403 (2001).
13. Blinov, B. B., Moehring, D. L., Duan, L. M. & Monroe, C. Observation of entanglement between a single trapped atom and a single photon. *Nature* **428**, 153–157 (2004).
14. Hald, J., Sørensen, J. L., Schori, C. & Polzik, E. S. Spin squeezed atoms: A macroscopic entangled ensemble created by light. *Phys. Rev. Lett.* **83**, 1319–1322 (1999).
15. Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* **409**, 490–493 (2001).
16. Phillips, D. F., Fleischhauer, A., Mair, A., Walsworth, R. L. & Lukin, M. D. Storage of light in atomic vapor. *Phys. Rev. Lett.* **86**, 783–786 (2001).
17. Schori, C., Julsgaard, B., Sørensen, J. L. & Polzik, E. S. Recording quantum properties of light in a long-lived atomic spin state: Towards quantum memory. *Phys. Rev. Lett.* **89**, 057903 (2002).
18. Julsgaard, B., Schori, C., Sørensen, J. L. & Polzik, E. S. Atomic spins as a storage medium for quantum fluctuations of light. *Quant. Inform. Comput.* **3** (special issue), 518–534 (2003).
19. Kuzmich, A. & Polzik, E. S. Atomic quantum state teleportation and swapping. *Phys. Rev. Lett.* **85**, 5639–5642 (2000).
20. Julsgaard, B., Sherson, J., Sørensen, J. L. & Polzik, E. S. Characterizing the spin state of an atomic ensemble using the magneto-optical resonance method. *J. Opt. B* **6**, 5–14 (2004).
21. Fiurášek, J., Cerf, N. J. & Polzik, E. S. Quantum cloning at the light-atoms interface: copying a coherent light state into two atomic quantum memories. *Phys. Rev. Lett.* (in the press); preprint at (<http://xxx.lanl.gov/abs/quant-ph/0404054>) (2004).
22. Hammerer, K., Mølmer, K., Polzik, E. S. & Cirac, J. I. Light-matter quantum interface. *Phys. Rev. A* (in the press); preprint at (<http://xxx.lanl.gov/abs/quant-ph/0312156>) (2004).

Supplementary Information accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We are grateful to N. Cerf and K. Hammerer for discussions. This research was funded by the Danish National Research Foundation, by EU grants QUICOV, COVAQIAL and CHIC, and by the project 'Research Center for Optics' of the Czech Ministry of Education. I.C. and E.S.P. acknowledge the hospitality of the Institute for Photonic Sciences, Barcelona, where part of this work was initiated.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to E. P. (polzik@nbi.dk).

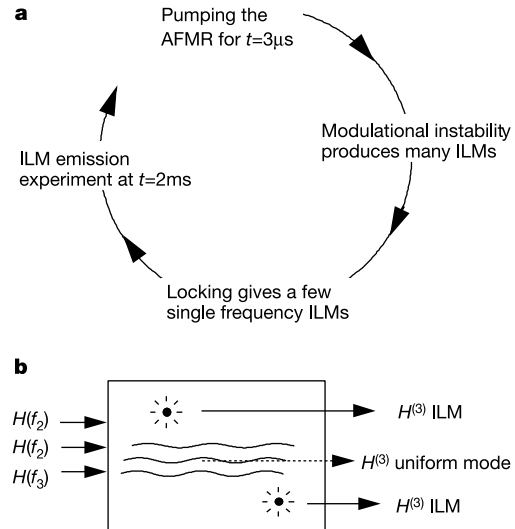
## Direct observation of the discrete character of intrinsic localized modes in an antiferromagnet

M. Sato & A. J. Sievers

Laboratory of Atomic and Solid State Physics and Cornell Center for Materials Research, Cornell University, Ithaca, New York 14850-2501, USA

In a strongly nonlinear discrete system, the spatial size of an excitation can become comparable to, and influenced by, the lattice spacing. Such intrinsic localized modes (ILMs)—also called 'discrete breathers' or 'lattice solitons'—are responsible for energy localization in the dynamics of discrete nonlinear lattices<sup>1–5</sup>. Their energy profiles resemble those of localized modes of defects in a harmonic lattice but, like solitons, they can move (although, unlike solitons, some energy is exchanged during collisions between them). The manipulation of these localized energy 'hotspots' has been achieved in systems as diverse as annular arrays of coupled Josephson junctions<sup>6,7</sup>, optical waveguide arrays<sup>8</sup>, two-dimensional nonlinear photonic crystals<sup>9</sup> and micromechanical cantilever arrays<sup>10</sup>. There is also some evidence for the existence of localized excitations in atomic lattices<sup>11–15</sup>, although individual ILMs have yet to be identified. Here we report the observation of countable localized excitations in an antiferromagnetic spin lattice by means of a nonlinear spectroscopic technique. This detection capability permits the properties of individual ILMs to be probed; the disappearance of each ILM registers as a step in the time-dependent signal, with the surprising result that the energy staircase of ILM excitations is uniquely defined.

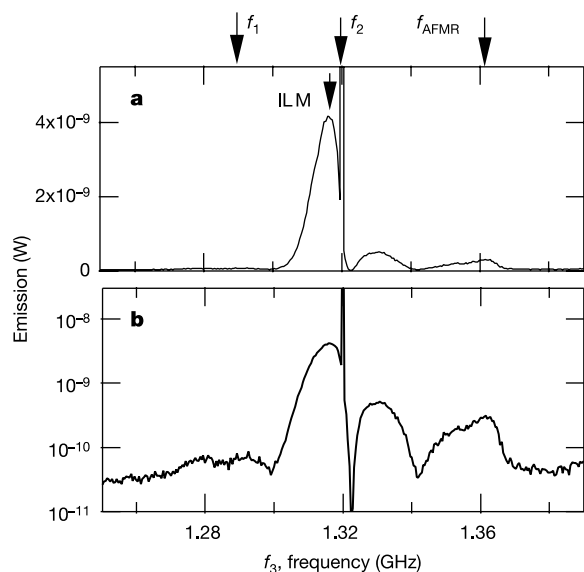
Rod-shaped samples of the quasi-one-dimensional antiferromagnet,  $(\text{C}_2\text{H}_5\text{NH}_3)_2\text{CuCl}_4$ , are used in this study<sup>13,16–18</sup> so that the antiferromagnetic resonance (AFMR) frequency occurs at the bottom of the spin-wave manifold. Mesoscale, high spin-precession amplitude, magnetic ILMs can then be produced in the gap below this band because of the soft nonlinearity of the spin lattice. As outlined in Fig. 1a, there are four fundamental sequential time steps required for these energy localization experiments: (1) a uniform spin wave mode of the biaxial antiferromagnet is driven to a large precession amplitude by an initial microwave pulse (frequency  $f_1$ ); (2) after an incubation period, the modulational instability of the large amplitude uniform mode takes hold, producing many ILMs in a broad frequency band<sup>19</sup>; (3) a few of these ILMs are then locked<sup>20,21</sup> to the continuous wave (c.w.) middle power source  $f_2$ ; and, finally, (4) the experiment of interest, where a c.w. low power source  $f_3$  is used to produce a mixing signal that depends on the number of ILMs in the lattice. After a spin-lattice relaxation time of  $T_1 = 1.5$  ms, the number of ILMs is determined by a quasi-steady state involving the  $f_2$  driver and the nonequilibrium AFMR<sup>13</sup>, and this number is expected to decrease as this frequency difference increases.



**Figure 1** Schematic diagrams of the experimental procedure and the nonlinear process. **a**, The chronological timing in the nonlinear energy magnetometer is shown for the production and measurement of intrinsic localized modes (ILMs). AFMR, antiferromagnetic resonance. **b**, Nonlinear mixing output for uniform excitations and for ILMs. Large filled circles represent ILMs. Although small in number because of locking, these strong nanoscale third order mixing elements dominate the output signal. Two inputs are from the locking driver field  $H(f_2)$  and the remainder is from the probe field  $H(f_3)$ , giving a maximum output at  $2f_2 - f_3$ . Each ILM generates a nonlinear field  $H^{(3)}$ . As the mixed signal power depends on the square of the total field, the square root of the power is proportional to the number of ILMs.

Because the number of spins in an ILM is too small to be seen in absorption, a nonlinear energy magnetometer has been developed, which automatically handles the four time steps for this low temperature ILM production and detection experiment. It relies on the third-order nonlinearity  $\chi^{(3)}$  of the antiferromagnet<sup>22–24</sup> to make observable in nonlinear emission the small number of ILMs that remain locked to the  $f_2$  driver, as illustrated in Fig. 1b. Nonlinear and linear excitations in the same sample give very different nonlinear responses: the signal produced by ILMs is enhanced, while that produced by the small precession amplitude extended plane wave states, which are nearly harmonic, is suppressed. This nonlinear discrimination feature of the instrument brings the few nanoscale ILMs out of the background of plane wave states for experimental exploration.

Presented in Fig. 2a is the power spectrum produced at  $f_3$ , where  $2f_2 - f_3 = f_{\text{det}}$ , and both  $f_3$  and  $f_{\text{det}}$ , the narrow band detector frequency, are scanned in tandem. Typical frequency positions for  $f_1$ ,  $f_2$  and the near-equilibrium AFMR are identified at the top of Fig. 2. The strong emission peak from the locked ILMs (see arrow) is observed shifted to slightly lower frequencies from the  $f_2$  locking oscillator (about 4 MHz), while a somewhat weaker peak is shifted up in frequency. (These sidebands represent a form of cross phase modulation for the ILMs.) The logarithmic plot shown in Fig. 2b displays more clearly the other weaker spectral features, as well as the excellent signal to noise level for these measurements. The weak peak observed at 1.362 GHz (see top right arrow) occurs when  $f_3$  matches the AFMR frequency in a resonant four wave mixing process at  $T \approx 2$  ms. This weak feature at long times is independent of whether or not the  $f_1$  pulse is applied and, hence, is a property of the homogeneous solid. When ILMs are produced, they encompass states originally associated with the spin wave band; however, because of the small number of ILMs studied (that is, only a few ILMs remain locked), 99.999% of the spins are in the low amplitude homogeneous state. Figure 2b demonstrates that although the AFMR involves many spins, off resonance it is a very weak nonlinear mixer.

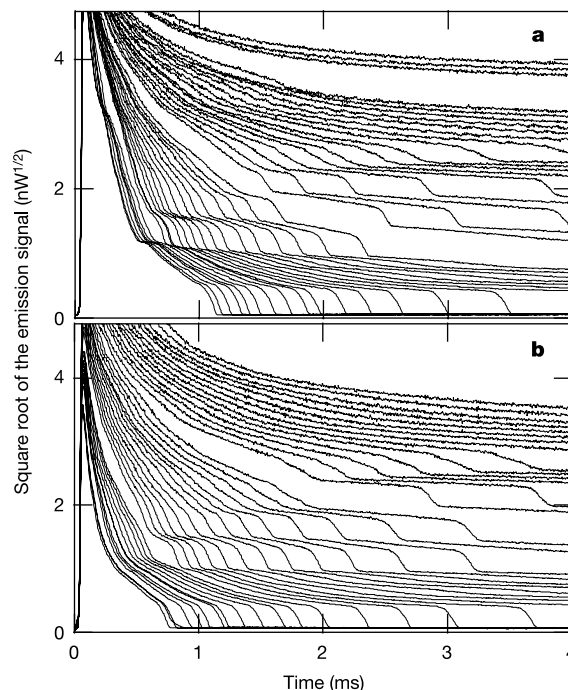


**Figure 2** Snapshot of the mixing spectrum versus the probe oscillator frequency. **a**, Mixing data taken at 2 ms after the 3- $\mu$ s-long, 52 W pulse at  $f_1 = 1.29$  GHz. Here  $f_2 = 1.32$  GHz at a c.w. power of 240 mW. The weak ( $\sim 1$  mW) probe oscillator of variable frequency  $f_3$  is scanned. A number of features are seen in the  $H^{(3)}$  emission. The strongest are associated with the disappearance of locked ILMs. The sample is immersed in 1.2 K liquid helium. **b**, Logarithmic plot of the same mixing data to bring out the weaker spectral features. The peak at 1.362 GHz comes from the resonant four wave mixing signal with the AFMR, and is also observed without the  $f_1$  pulse.

To investigate the time dependence of the mixing signal from the ILMs,  $f_3$  is now set to the frequency of the emission maximum in Fig. 2a. Typical time-dependent emission results are shown in Fig. 3. In Fig. 3a the different traces correspond to a fixed  $f_2$  frequency with its power setting varied incrementally, whereas in Fig. 3b the  $f_2$  power is fixed but its frequency is changed. The general trends in both panels are similar. Over some range of parameter space, a smooth decrease in the emission signal with time occurs because the AFMR is still far from its equilibrium value and controls the mixing signal, while for other ranges of parameters, when the AFMR is near its equilibrium value and the ILM component dominates, steps appear in the time-dependent emission. Although the appearance of steps as individual ILMs become unlocked is not unexpected, the uniformity is completely unexpected. The square root of the emission power is displayed in Fig. 3 to bring out the similarity of the step structure observed for the different parameters. The zero level corresponds to no emission, and the height of the lowest step is seen to depend only weakly on time, and hence in Fig. 3a on the  $f_2$  power level and in Fig. 3b on the  $f_2$  frequency. These surprisingly weak dependences of the step height carry over to the higher steps as well. Experiments with different samples and different coax-sample coupling coils always produce such steps, although the step magnitude for different combinations of components can vary by up to 50%.

Another intriguing pattern displayed in both parts of Fig. 3 is the absence of signal in certain well-defined regions. These time-dependent gaps may be a feature produced by interference between the nonresonant four wave mixing signal from the wing of the antiferromagnetic resonance and that produced by the resonant four wave mixing signal from the ILMs.

We find that the traces in Fig. 3a and b can be decomposed into an exponential time-dependent background plus steps of relatively fixed heights. If step heights of different traces are counted at a particular time delay of about 2 ms, then for the incremental power scan in Fig. 3a the mean value is  $0.29 \pm 0.04$  (nW) $^{1/2}$ , whereas for the frequency scan in Fig. 3b the mean value is  $0.30 \pm 0.01$  (nW) $^{1/2}$ .



**Figure 3** Experimental demonstration of ILM energy quantization. **a**, Square root of the time-dependent emission output as a function of the c.w.  $f_2$  power level. The 2.3% increments between curves vary the  $f_2$  power from 34.7 to 87.1 mW. Superimposed on the smooth time-dependent signal are steps in many of the traces. When an ILM disappears, a step is recorded. **b**, Square root of the time-dependent emission output as a function of the  $f_2$  frequency. The frequency is scanned from 1.33 to 1.34 GHz at a power level of 55 mW so that  $0.01/(f_{\text{AFMR}} - f_2) \approx 33\%$ .

For this latter case, the  $\Delta f$  gap between the ILM frequency and the AFMR frequency has been varied by 33% from the top trace to the lowest trace. ILM eigenvector calculations<sup>25</sup> indicate that the step height should have changed by at least 20%, but such a variation is not observed. Classical explanations fail to account for these nearly constant step heights in the nonlinear emission data.

The change in four wave mixing magnetization associated with a step—that is, the disappearance of an individual ILM—can be estimated by combining a few basic ideas together with standard homogeneous four wave mixing results to find that the emitted power varies as the fourth power of the magnitude of the transverse magnetization at  $f_2$ ; hence, the square root of the power varies as the dynamical energy. A comparison of the measured step height with the calculated one is used to find the volume fraction of an ILM in this crystal. We find that the number of spins in an ILM is  $N_{\text{spin}} = 8 \times 10^{14}$ , and that the energy associated with a single step is  $\Delta E = 1.5 \times 10^{-12}$  J, so the energy per  $\text{Cu}^{2+}$  spin  $\Delta E/N_{\text{spin}} = 1.9 \times 10^{-27}$  J. The rigidity of this emission step size against variations in experimental parameters demonstrates that the energy of the observed ILM excitation is surprisingly well defined. □

Received 20 June; accepted 13 September 2004; doi:10.1038/nature03038.

1. Kiselev, S. A., Bickham, S. R. & Sievers, A. J. Properties of intrinsic localized modes in one-dimensional lattices. *Comments Condens. Matter Phys.* **17**, 135–173 (1995).
2. Flach, S. & Willis, C. R. Discrete breathers. *Phys. Rep.* **295**, 182–264 (1998).
3. Lai, R. & Sievers, A. J. Nonlinear nanoscale localization of magnetic excitations in atomic lattices. *Phys. Rep.* **314**, 147–236 (1999).
4. Campbell, D. K., Flach, S. & Kivshar, Y. S. Localizing energy through nonlinearity and discreteness. *Phys. Today* **57**, 43–49 (2004).
5. Flach, S. in *Energy Localisation and Transfer* (eds Dauxois, T., Litvak-Hinzenon, A., MacKay, R. & Spanoudaki, A.) 1–71 (World Scientific, London, 2004).
6. Trias, E., Mazo, J. J. & Orlando, T. P. Discrete breathers in nonlinear lattices: Experimental detection in a Josephson array. *Phys. Rev. Lett.* **84**, 741–744 (2000).
7. Binder, P., Abaimov, D., Ustinov, A. V., Flach, S. & Zolotarev, Y. Observation of breathers in Josephson ladders. *Phys. Rev. Lett.* **84**, 745–748 (2000).



8. Eisenberg, H. S., Silberberg, Y., Morandotti, R., Boyd, A. R. & Aitchison, J. S. Discrete spatial optical solitons in waveguide arrays. *Phys. Rev. Lett.* **81**, 3383–3386 (1998).
9. Fleischer, J. W., Segev, M., Efremidis, N. K. & Christodoulides, D. N. Observation of two-dimensional discrete solitons in optically induced nonlinear photonic lattices. *Nature* **422**, 147–150 (2003).
10. Sato, M., Hubbard, B. E., Sievers, A. J., Ilic, B. & Craighead, H. G. Optical manipulation of intrinsic localized vibrational energy in cantilever arrays. *Europhys. Lett.* **66**, 318–323 (2004).
11. Swanson, B. I. *et al.* Observation of intrinsically localized modes in a discrete low dimensional material. *Phys. Rev. Lett.* **82**, 3288–3291 (1999).
12. Fehske, H., Kinatader, M., Wellein, G. & Bishop, A. R. Quantum lattice effects in mixed-valence transition-metal chain complexes. *Phys. Rev. B* **63**, 245121 (2001).
13. Schwarz, U. T., English, L. Q. & Sievers, A. J. Experimental generation and observation of intrinsic localized spin wave modes in an antiferromagnet. *Phys. Rev. Lett.* **83**, 223–227 (1999).
14. Xie, A., van der Meer, L., Hoff, W. & Austin, R. H. Long-lived Amide I vibrational modes in myoglobin. *Phys. Rev. Lett.* **84**, 5435–5438 (2000).
15. Markovich, T., Polturak, E., Bossy, J. & Farhi, E. Observation of a new excitation in bcc He-4 by inelastic neutron scattering. *Phys. Rev. Lett.* **88**, 195301 (2002).
16. Chikamatsu, M., Tanaka, M. & Yamazaki, H. Effect of magnetic dipolar interaction on antiferromagnetic resonance in  $(\text{C}_2\text{H}_5\text{NH}_3)_2\text{CuCl}_4$ . *J. Phys. Soc. Jpn* **50**, 2876–2883 (1981).
17. de Jongh, L. J. in *Physics and Chemistry of Materials with Low-Dimensional Structures* (ed. Lévy, F.) 1–51 (Kluwer Academic, Dordrecht, 1990).
18. Sato, M., English, L. Q., Hubbard, B. E. & Sievers, A. J. Influence of sample shape on the production of intrinsic localized modes in an antiferromagnetic lattice. *J. Appl. Phys.* **91**, 8676–8678 (2002).
19. English, L. Q. *Studies of Intrinsic Localized Spin-Wave Modes in Antiferromagnetic Crystals*. Ph.D. thesis, Cornell Univ. (2003).
20. Rössler, T. & Page, J. B. Intrinsic localized modes in driven anharmonic lattices with realistic potentials. *Phys. Lett. A* **204**, 418–426 (1995).
21. Sato, M. *et al.* Observation of locked intrinsic localized vibrational modes in a micromechanical oscillator array. *Phys. Rev. Lett.* **90**, 044102 (2003).
22. Shen, Y. R. *The Principles of Nonlinear Optics* (John Wiley & Sons, New York, 1984).
23. Almeida, N. S. & Mills, D. L. Nonlinear infrared response of antiferromagnets. *Phys. Rev. B* **36**, 2015–2023 (1987).
24. Lim, S.-C., Osman, J. & Tilley, D. R. Calculation of nonlinear magnetic susceptibility tensors for a uniaxial antiferromagnet. *J. Phys. D* **33**, 2899–2910 (2000).
25. Lai, R. & Sievers, A. J. Identification of an intrinsic localized spin wave resonance in antiferromagnetic chains with single-ion easy-plane anisotropy. *Phys. Rev. B* **55**, 11937–11940 (1997).

**Acknowledgements** We acknowledge conversations with P. Brouwer, L. English, B. Hubbard, D. Mills, J. Page, J. Parpia, J. Sethna and R. Silsbee. This work was supported by NSF-DMR and by the Cornell Center for Materials Research.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to A.J.S. (siewers@ccmr.cornell.edu).

## Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors

Kenji Nomura<sup>1</sup>, Hiromichi Ohta<sup>1</sup>, Akihiro Takagi<sup>2</sup>, Toshio Kamiya<sup>1,2</sup>, Masahiro Hirano<sup>1</sup> & Hideo Hosono<sup>1,2,3</sup>

<sup>1</sup>ERATO-SORST, JST, in Frontier Collaborative Research Center, Mail Box S2-13, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

<sup>2</sup>Materials and Structures Laboratory, Mail Box R3-1, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

<sup>3</sup>Frontier Collaborative Research Center, Mail Box S2-13, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

Transparent electronic devices formed on flexible substrates are expected to meet emerging technological demands where silicon-based electronics cannot provide a solution. Examples of active flexible applications include paper displays and wearable computers<sup>1</sup>. So far, mainly flexible devices based on hydrogenated amorphous silicon (a-Si:H)<sup>2–5</sup> and organic semiconductors<sup>2,6–10</sup> have been investigated. However, the performance of these devices has been insufficient for use as transistors in practical

computers and current-driven organic light-emitting diode displays. Fabricating high-performance devices is challenging, owing to a trade-off between processing temperature and device performance. Here, we propose to solve this problem by using a novel semiconducting material—namely, a transparent amorphous oxide semiconductor from the In-Ga-Zn-O system (a-IGZO)—for the active channel in transparent thin-film transistors (TTFTs). The a-IGZO is deposited on polyethylene terephthalate at room temperature and exhibits Hall effect mobilities exceeding  $10\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ , which is an order of magnitude larger than for hydrogenated amorphous silicon. TTFTs fabricated on polyethylene terephthalate sheets exhibit saturation mobilities of  $6\text{--}9\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ , and device characteristics are stable during repetitive bending of the TTFT sheet.

Thin-film transistors (TFTs) are fundamental building blocks for state-of-the-art microelectronics, such as flat-panel displays and system-on-glass<sup>11,12</sup>. Furthermore, the fabrication of low-temperature TFTs will allow flexible large-area electronic devices to be developed. These devices are flexible, lightweight, shock resistant and potentially affordable—properties that are necessary for large, economic, high-resolution displays, wearable computers and paper displays<sup>1</sup>. Further, when combined with ‘transparent circuit technology’<sup>13–17</sup>, TFTs can integrate display functions even on the windscreens of cars.

Organic semiconductors and hydrogenated amorphous silicon (a-Si:H) have been extensively investigated for flexible electronics, and have demonstrated the ability to be fabricated into flexible solar cells and TFTs<sup>2–10</sup>. However, device performance is limited by the low mobilities of the channel materials (field effect mobilities,  $\mu_{\text{FE}}$ , are  $\sim 1\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$  for a-Si:H,  $\sim 2.7\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$  for a pentacene single crystal, and  $\sim 1.5\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$  for a pentacene thin film<sup>10</sup>). In addition, Si-based devices are of less interest for transparent circuits because they are not transparent, owing to the small bandgap<sup>18</sup>.

Amorphous semiconductors are preferred over polycrystalline ones for active layers from the viewpoints of processing temperature and uniformity of device characteristics. However, the carrier mobility of a-Si:H is lower by two or three orders of magnitude than that of single-crystalline Si ( $\sim 200\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$  for carrier concentration  $\sim 10^{19}\text{ cm}^{-3}$ ). The mobility of a-Si:H is limited to  $\sim 1\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ , as carrier transport is controlled by hopping between localized tail-states and band conduction is not achieved. The low mobility is associated with the intrinsic nature of the chemical bonding (Fig. 1a): average carrier transport paths in covalent semiconductors such as a-Si:H consist of  $sp^3$  orbitals with strong directivity and, therefore, the bond angle fluctuation significantly alters the electronic levels, leading to somewhat high-density deep tail-states.

In contrast, degenerate band conduction and large mobility ( $>10\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ ) are possible in amorphous oxide semiconductors (AOSs) containing post-transition-metal cations<sup>19,20</sup>. These features are completely different from those of the covalent semiconductors. Figure 1b illustrates the carrier transport paths (that is, the wavefunction of the conduction band bottom) in AOSs. The bottom of the conduction band in the oxide semiconductors that has high ionicity is primarily composed of spatially spread metal  $ns$  orbitals with isotropic shape (here  $n$  is the principal quantum number), and direct overlap among the neighbouring metal  $ns$  orbitals is possible. The magnitude of this overlap is insensitive to distorted metal–oxygen–metal (M–O–M) chemical bonds that intrinsically exist in amorphous materials<sup>21,22</sup>. Therefore, AOSs exhibit Hall-effect mobilities similar to those of the corresponding crystalline phase, even if they are formed at room temperature. These carrier transport properties are unique to oxide semiconductors, and are not seen in covalent amorphous semiconductors such as a-Si:H.

Here we report room-temperature fabrication and performance of flexible TTFTs fabricated using a-IGZO as an active n-channel

layer. Films of a-IGZO were prepared by pulsed laser deposition with a KrF excimer laser, using a polycrystalline InGaZnO<sub>4</sub> target at room temperature in an oxygen atmosphere (oxygen pressure  $P_{O_2}$ ). The chemical composition of the obtained films measured by X-ray fluorescence spectroscopy was In:Ga:Zn = 1.1:1.1:0.9 (in atomic ratio).

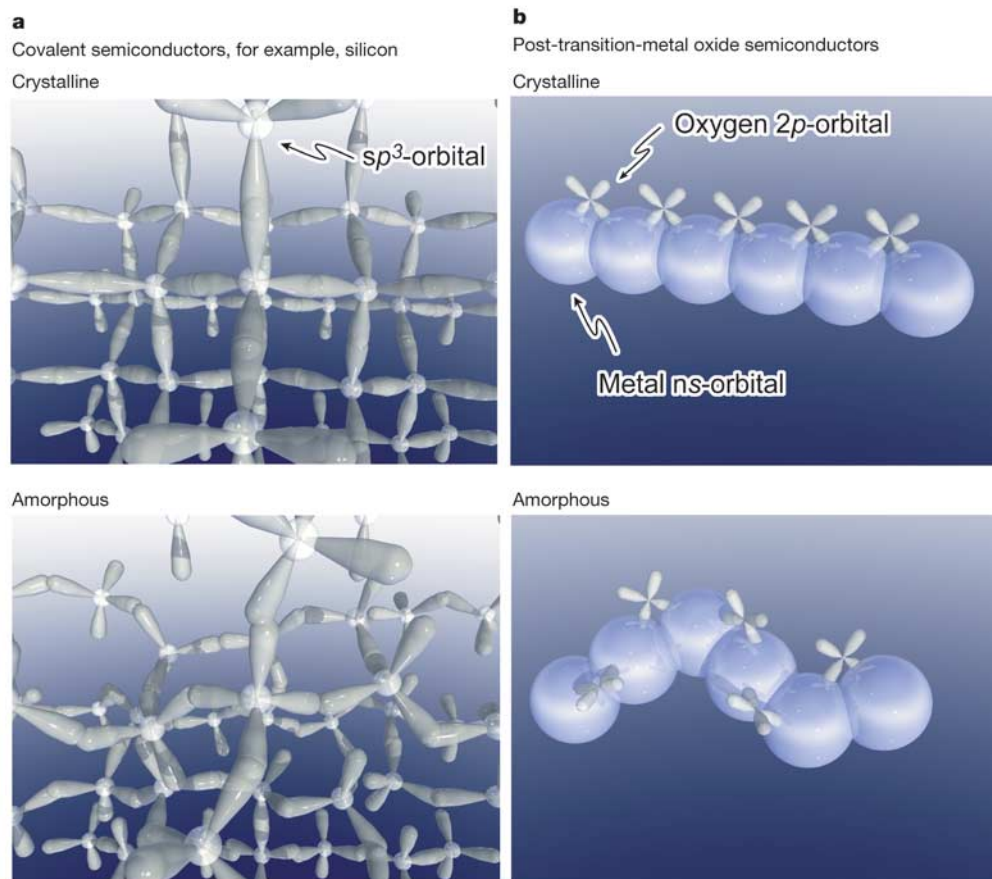
Figure 2a shows an X-ray diffraction pattern of an a-IGZO film deposited on a glass substrate. The film is amorphous and optically transparent in the entire visible and near-infrared regions (wavelength  $\lambda = 390\text{--}3,200\text{ nm}$ ), as shown in Fig. 2a inset. The optical transmittance is greater than 80%, including the reflection associated with the film and glass substrate. The optical bandgap ( $E_g$ ) estimated from the Tauc' plot is  $\sim 3.0\text{ eV}$ , which is similar to that of the crystalline phase ( $\sim 3.4\text{ eV}$ ). The electrical conductivities at room temperature are  $<10^{-5}\text{ S cm}^{-1}$  when the films are deposited at  $P_{O_2} > 6\text{ Pa}$ . The values correspond to carrier concentrations of  $<10^{14}\text{ cm}^{-3}$  if an electron mobility of  $1\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$  is assumed. The sign of the Seebeck coefficients obtained from thermopower measurements is negative, indicating that a-IGZO is an n-type semiconductor.

Figure 2b shows the room-temperature Hall mobility of a-IGZO films as a function of carrier concentration. This carrier concentration varies from  $<10^{14}$  to  $10^{20}\text{ cm}^{-3}$  when  $P_{O_2}$  is varied from  $\sim 7$  to  $0.1\text{ Pa}$ . The data obtained on single-crystalline InGaO<sub>3</sub>(ZnO)<sub>5</sub> (c-IGZO) films are shown for comparison. We have reported<sup>23</sup> that carrier transport in c-IGZO is governed by percolation conduction over the distribution of potential barriers around the conduction

band edge; these potential barriers are formed owing to random distribution of Ga<sup>3+</sup> and Zn<sup>2+</sup> ions in the crystal structure. The potential barriers are overcome when carrier concentration exceeds  $3 \times 10^{18}\text{ cm}^{-3}$ , and therefore the Hall mobility increases as the carrier concentration increases, and larger Hall mobilities ( $>10\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ ) are obtained at carrier concentrations  $>10^{18}\text{ cm}^{-3}$ . Similar behaviours are also observed in a-IGZO, which would result from a similar mechanism associated with the random amorphous structure.

The amorphous phase is thermally stable up to  $\sim 500^\circ\text{C}$  in air. Other AOSs such as a-ITO, and also crystalline ZnO (an amorphous phase of ZnO has yet been reported), have high-density carriers even in as-deposited states, and are difficult to make into devices with controlled characteristics. Thus it is vital to choose a material in which carrier concentration can be controlled at a low level, for example,  $<10^{14}\text{ cm}^{-3}$ , in order to achieve a low off current and large on-to-off current ratios<sup>24</sup>. Incorporating Ga ions would be important in a-IGZO for suppressing carrier generation via oxygen vacancy formation, because the Ga ion forms stronger chemical bonds with oxygen than Zn and In ions.

We fabricated top-gate flexible TFTs using the a-IGZO film as an n-channel active layer on 200- $\mu\text{m}$ -thick polyethylene terephthalate (PET) films (Fig. 3a). Source, drain, gate contacts and a gate insulator were defined by standard photolithography and lift-off techniques. A  $\sim 140\text{-nm}$ -thick Y<sub>2</sub>O<sub>3</sub> layer was chosen for the gate insulator and ITO (Sn:10%) was used for source, drain and gate transparent electrodes. These layers were deposited by pulsed laser



**Figure 1** Schematic orbital drawings for the carrier transport paths (that is, conduction band bottoms) in crystalline and amorphous semiconductors. **a**, Covalent semiconductors have carrier transport paths composed of strongly directive  $sp^3$  orbitals, so structural randomness greatly degrades the magnitude of bond overlap, that is, carrier mobility. Note that the orbitals shown are illustrative, and do not show exact wavefunctions.

**b**, Amorphous oxide semiconductors composed of post-transition-metal cations. Spheres denote metal  $s$  orbitals. The contribution of oxygen  $2p$  orbitals is small. Direct overlap between neighbouring metal  $s$  orbitals is rather large, and is not significantly affected even in an amorphous structure.

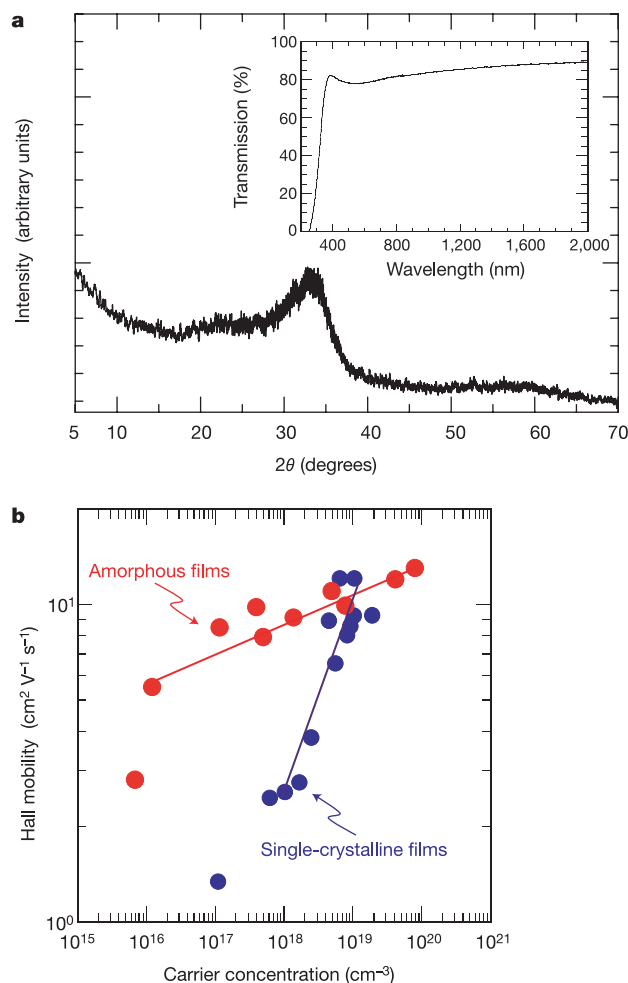
deposition at room temperature using  $\text{Y}_2\text{O}_3$  and ITO ceramic targets. The measured dielectric constant of the  $\text{Y}_2\text{O}_3$  gate insulator was  $\sim 16\epsilon_0$  (where  $\epsilon_0$  is the dielectric constant of vacuum), which is close to that reported for crystalline  $\text{Y}_2\text{O}_3$  ( $\sim 18\epsilon_0$ )<sup>25</sup>. The channel length and width were 50  $\mu\text{m}$  and 200  $\mu\text{m}$ , respectively (Fig. 3c).

The performances of the flexible TTFTs were measured in air at room temperature. Figure 4a shows typical source-to-drain current ( $I_{\text{DS}}$ )–voltage ( $V_{\text{DS}}$ ) characteristics of a virgin device. The current  $I_{\text{DS}}$  markedly increases as  $V_{\text{DS}}$  increases at a positive gate bias ( $V_{\text{GS}}$ ), indicating that the channel is n-type. The  $I_{\text{DS}}$  reaches  $\sim 0.02$  mA at a  $V_{\text{GS}}$  of 5 V. The  $I_{\text{DS}}-V_{\text{DS}}$  characteristics exhibit a clear pinch-off and current saturation, confirming that the TTFT operation follows the standard field-effect transistor theory. The saturation mobility ( $\mu_{\text{sat}}$ ) is obtained from the  $I_{\text{DS}}-V_{\text{DS}}$  curves in the saturation region using the equation  $I_{\text{DS,sat}} = (C_i \mu_{\text{sat}} W/2L)(V_{\text{GS}} - V_{\text{T}})^2$ , where  $C_i$ ,  $V_{\text{T}}$ ,  $W$  and  $L$  denote the gate capacitance, threshold gate voltage, channel width and length, respectively. The estimated  $\mu_{\text{sat}}$  value is  $\sim 8.3 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ , much larger than those obtained in organic and a-Si:H TFTs (the  $\mu_{\text{sat}}$  value ranges from  $\sim 6$  to  $\sim 9 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  in

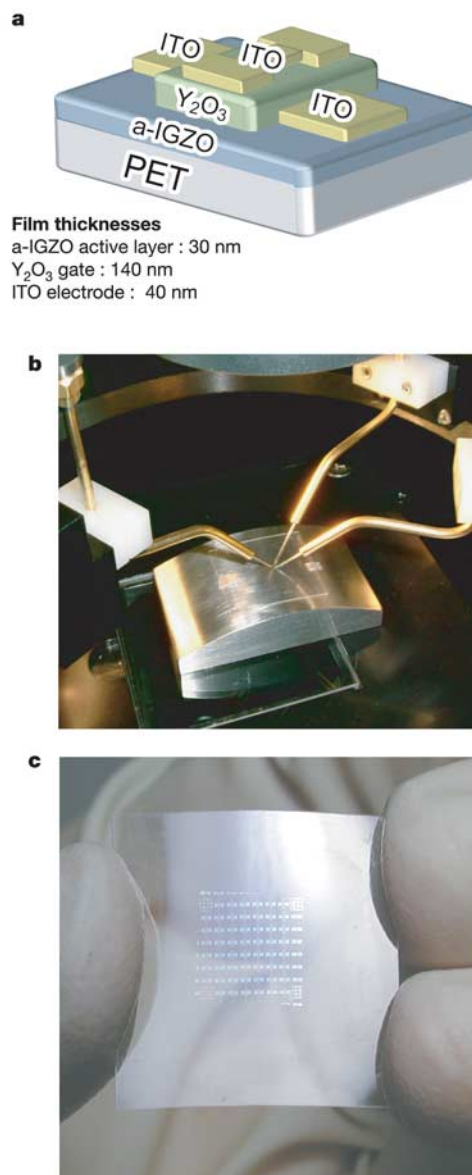
the devices measured). Similarly, a field-effect mobility ( $\mu_{\text{FE}}$ ) estimated from the linear region in the  $I_{\text{DS}}-V_{\text{DS}}$  curve (using  $I_{\text{DS}} = (C_i \mu_{\text{FE}} W/L)(V_{\text{GS}} - V_{\text{T}})V_{\text{DS}}$ ) is  $\sim 5.6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at  $V_{\text{DS}} = 2.4$  V, which agrees roughly with the  $\mu_{\text{sat}}$  value.

The transfer characteristic (Fig. 4b) shows that a low off-current, of the order of  $10^{-7}$  A, and an on-to-off current ratio  $\sim 10^3$  are obtained. The threshold gate voltage is positive ( $V_{\text{T}} \approx +1.6$  V), showing that the TTFT operates in the enhancement mode (normally-off characteristics). The gate leak current is lower by several orders of magnitude than the source-to-drain current, which guarantees that the TFT characteristics are unaffected by the gate leak current. Small hystereses were observed in the  $I_{\text{DS}}-V_{\text{GS}}$  curves with negative shifts in gate bias of 0.1–0.5 V at a scan speed of  $0.5 \text{ V s}^{-1}$ , which would be due to interface states.

Next, bending effects on the TTFT characteristics were examined.

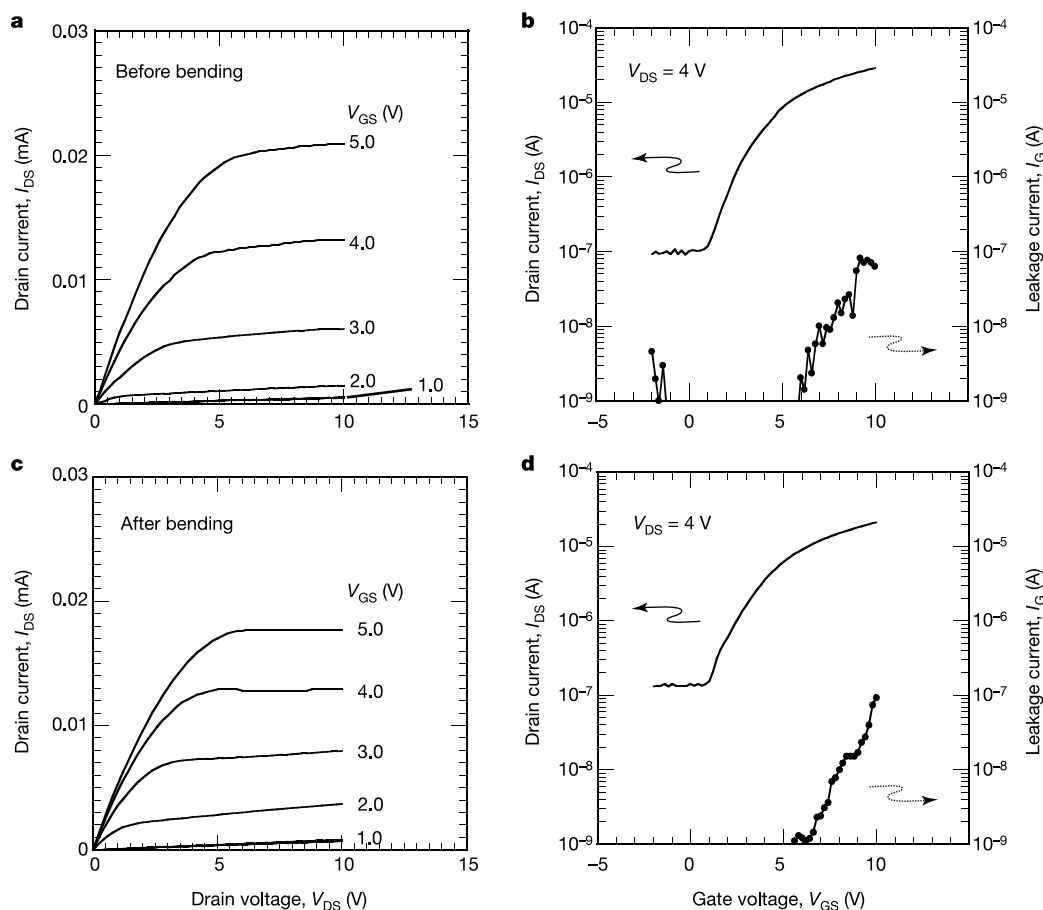


**Figure 2** Amorphous IGZO films. **a**, Glance angle X-ray diffraction pattern of the a-IGZO film deposited on a silica glass substrate at room temperature. No sharp peak is observed.  $\text{CuK}_{\alpha 1}$  radiation was used. Inset is the optical transmission spectrum of the film. The measured bandgap ( $E_g$ ) is  $\sim 3.0$  eV. **b**, Relationship between room-temperature Hall mobility and carrier concentration for a-IGZO films. Data on single-crystalline  $\text{InGaO}_3(\text{ZnO})_5$  films are shown for comparison. Electron mobility strongly depends on carrier concentration, and exceeds  $10 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at carrier concentrations greater than  $10^{18} \text{ cm}^{-3}$ , owing to potential distribution in the vicinity of the conduction band bottom (see ref. 23 for details).



**Figure 3** Flexible TTFTs. **a**, Structure of TTFT fabricated on a plastic sheet. **b**, A photograph of the flexible TTFT sheet bent at  $R = 30$  mm. The TTFT sheet is fully transparent in the visible light region. **c**, A photograph of the flexible TTFT sheet. The transparent TFT devices are made visible by adjusting the angle of the illumination.





**Figure 4** Typical TFT characteristics before and after bending. **a**,  $I_{DS}$ – $V_{DS}$  and **b**,  $I_{DS}$ – $V_{GS}$  characteristics before bending. The TFT operates in the enhanced mode with a threshold voltage of  $\sim 1.6$  V. The saturation mobility is  $\sim 8.3$  cm<sup>2</sup> V<sup>−1</sup> s<sup>−1</sup>. On-to-off

current ratio is  $\sim 10^3$ . **c**,  $I_{DS}$ – $V_{DS}$  and **d**,  $I_{DS}$ – $V_{GS}$  characteristics after bending. The device was bent at  $R = 30$  mm.

The TTFT sheet was bent into a curve with a surface curvature radius ( $R$ ) of 30 mm (corresponding to a tensile strain of  $\sim 0.3\%$  in the TTFTs), as shown in Fig. 3b. The TTFT after bending maintained good characteristics, such as  $\mu_{\text{sat}} \approx 7$  cm<sup>2</sup> V<sup>−1</sup> s<sup>−1</sup> and an on-to-off current ratio of  $\sim 10^3$  (Fig. 4c, d). We stress that the TTFT performance is almost unaffected by bending, although a slight decrease is observed in the saturation current. After the initial bending, the TFT characteristics are stable and reproducible during and after repetitive bending. The TTFT is stable at temperatures up to 120 °C, but becomes inoperative at higher temperatures, probably owing to the softening of the PET substrate.

The present study demonstrates the room-temperature fabrication and operation of flexible TTFTs based on an amorphous oxide semiconductor, a-IGZO. TTFTs were fabricated on inexpensive polymer films and displayed good performance—such as saturation mobilities of  $\sim 6$ – $9$  cm<sup>2</sup> V<sup>−1</sup> s<sup>−1</sup>, a low leak current of  $\sim 10^{-10}$  A, and an on-to-off ratio of  $\sim 10^3$ —even during and after bending. We used pulsed laser deposition to form the active a-IGZO layer in this study, but a sputtering or metal-organic chemical vapour deposition (MOCVD) method can be used for large-area uniform deposition and mass production, as demonstrated for window electrodes of solar cells and flat-panel displays.

These achievements imply that transparent amorphous oxide semiconductors have the potential to overtake a-Si:H, and are promising materials for transparent flexible electronics. Furthermore, flexible TTFTs may be integrated with other already-developed devices that use a p-type amorphous oxide semiconduc-

tor and p–n junction diodes fabricated at room temperature<sup>26</sup>; this would extend the possibilities of flexible transparent electronic circuits. □

Received 4 July; accepted 5 October 2004; doi:10.1038/nature03090.

- Huitema, H. E. A. *et al.* Plastic transistors in active-matrix displays. *Nature* **414**, 599 (2001).
- Kagan, C. R. & Andry, P. (eds) *Thin Film Transistors* (Marcel Dekker, New York, 2003).
- Yang, C.-S., Smith, L. L., Arthur, C. B. & Parsons, G. N. Stability of low-temperature amorphous silicon thin film transistors formed on glass and transparent plastic substrates. *J. Vac. Sci. Technol. B* **18**, 683–689 (2000).
- Carey, P. G., Smith, P. M., Theiss, S. D. & Wickboldt, P. Polysilicon thin film transistors fabricated on low temperature plastic substrates. *J. Vac. Sci. Technol. A* **17**, 1946–1949 (2000).
- Wang, S. C. *et al.* Device transfer technology by backside etching for poly-Si thin-film transistors on glass/plastic substrate. *Jpn. J. Appl. Phys.* **42**, 1044–1046 (2003).
- Shaw, J. M. & Seidler, P. F. Organic electronics: Introduction. *IBM J. Res. Dev.* **45**, 3–9 (2001).
- Dimitrakopoulos, C. D. & Malenfant, P. R. L. Organic thin film transistors for large area electronics. *Adv. Mater.* **14**, 99–117 (2002).
- Choi, H. Y., Kim, S. H. & Jang, J. Self-organized organic thin-film transistors on plastic. *Adv. Mater.* **16**, 732–736 (2004).
- Lee, J. H. *et al.* Pentacene thin film transistors fabricated on plastic substrates. *Synth. Metals* **139**, 445–451 (2003).
- Dimitrakopoulos, C. D. & Mascaró, D. J. Organic thin-film transistors: A review of recent advances. *IBM J. Res. Dev.* **45**, 11–27 (2001).
- Taur, Y. & Ning, H. T. *Fundamentals of Modern VLSI Devices* (Cambridge Univ. Press, New York, 1988).
- Kuo, Y. (ed.) *Thin Film Transistors: Materials and Processes* (Kluwer Academic, Dordrecht, 2004).
- Thomas, G. Invisible circuits. *Nature* **389**, 907–908 (1997).
- Kawazoe, H. *et al.* P-type electrical conduction in transparent thin films of CuAlO<sub>2</sub>. *Nature* **389**, 939–942 (1997).
- Ohta, H. & Hosono, H. Transparent oxide electronics. *Mater. Today* **7**, 42–51 (2004).
- Wager, J. F. Transparent electronics. *Science* **300**, 1245–1246 (2003).
- Hosono, H. Built-in nanostructures in transparent oxides for novel photonic and electronic functions materials. *Int. J. Appl. Ceram. Technol.* **1**, 106–118 (2004).
- Madelung, O. (ed.) *Technology and Application of Amorphous Silicon* (Springer, Berlin, 2000).

19. Hosono, H. *et al.* Working hypothesis to explore novel wide band gap electrically conducting amorphous oxides and examples. *J. Non-Cryst. Solids* **198–200**, 165–169 (1996).
20. Orita, M. *et al.* Amorphous transparent conductive oxide  $\text{InGaO}_3(\text{ZnO})_m$  ( $m \leq 4$ ): a Zn 4s conductor. *Phil. Mag. B* **81**, 501–515 (2001).
21. Mott, N. F. Silicon dioxide and the chalcogenide semiconductors; similarities and differences. *Adv. Phys.* **26**, 363–391 (1977).
22. Narushima, S. *et al.* Electronic structure and transport properties in the transparent amorphous oxide semiconductor  $2\text{CdO-GeO}_2$ . *Phys. Rev. B* **66**, 35203 (2002).
23. Nomura, K. *et al.* Carrier transport in transparent oxide semiconductor with intrinsic structural randomness probed using single-crystalline  $\text{InGaO}_3(\text{ZnO})_5$  films. *Appl. Phys. Lett.* **85**, 1993–1995 (2004).
24. Nomura, K. *et al.* Thin film transistor fabricated in single-crystalline transparent oxide semiconductor. *Science* **300**, 1269–1272 (2003).
25. Kwo, J. *et al.* Properties of high  $k$  gate dielectrics  $\text{Gd}_2\text{O}_3$  and  $\text{Y}_2\text{O}_3$  for Si. *J. Appl. Phys.* **89**, 3920–3927 (2001).
26. Narushima, S. *et al.* P-type amorphous oxide semiconductor,  $\text{ZnRh}_2\text{O}_4$ , and room temperature fabrication of amorphous oxide P-N hetero-junction diodes. *Adv. Mater.* **15**, 1409–1413 (2003).

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to H.H. (hosono@msl.titech.ac.jp).

## Equilibrium cluster formation in concentrated protein solutions and colloids

Anna Stradner<sup>1</sup>, Helen Sedgwick<sup>2</sup>, Frédéric Cardinaux<sup>1</sup>, Wilson C. K. Poon<sup>2</sup>, Stefan U. Egelhaaf<sup>2,3</sup> & Peter Schurtenberger<sup>1</sup>

<sup>1</sup>Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

<sup>2</sup>School of Physics and COSMIC, The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK

<sup>3</sup>School of Chemistry, The University of Edinburgh, West Mains Road, Edinburgh EH9 3JJ, UK

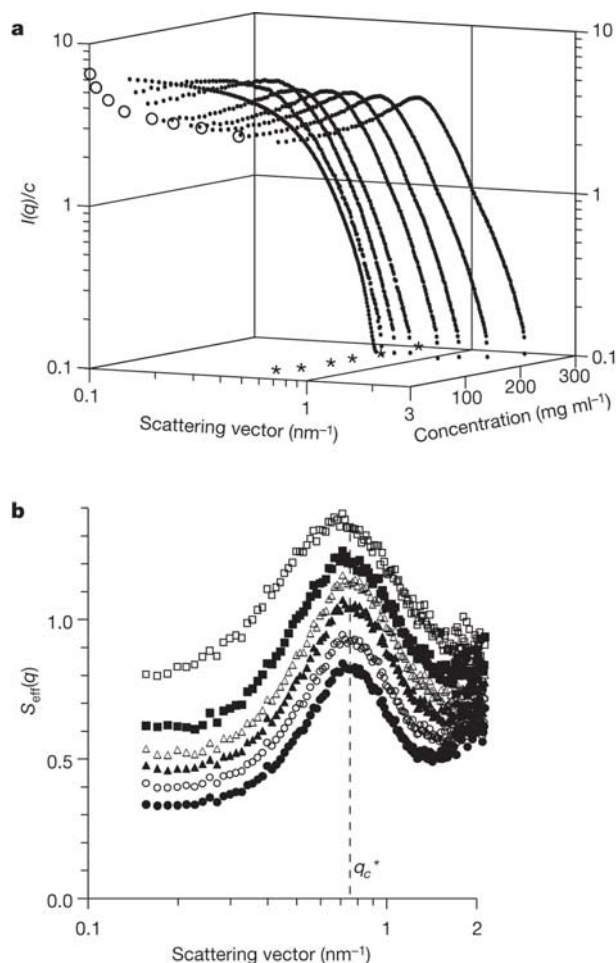
Controlling interparticle interactions, aggregation and cluster formation is of central importance in a number of areas, ranging from cluster formation in various disease processes to protein crystallography and the production of photonic crystals. Recent developments in the description of the interaction of colloidal particles with short-range attractive potentials have led to interesting findings including metastable liquid–liquid phase separation and the formation of dynamically arrested states (such as the existence of attractive and repulsive glasses, and transient gels)<sup>1–7</sup>. The emerging glass paradigm has been successfully applied to complex soft-matter systems, such as colloid–polymer systems<sup>8</sup> and concentrated protein solutions<sup>9</sup>. However, intriguing problems like the frequent occurrence of cluster phases remain<sup>10–13</sup>. Here we report small-angle scattering and confocal microscopy investigations of two model systems: protein solutions and colloid–polymer mixtures. We demonstrate that in both systems, a combination of short-range attraction and long-range repulsion results in the formation of small equilibrium clusters. We discuss the relevance of this finding for nucleation processes during protein crystallization, protein or DNA self-assembly and the previously observed formation of cluster and gel phases in colloidal suspensions<sup>12–17</sup>.

A number of globular proteins have been shown to exhibit the major characteristics of colloids that interact via a short-range attractive potential. At high ionic strength, where the salt screens electrostatic repulsions, these short-range attractions increasingly dominate with decreasing temperature. This leads to a metastable liquid–liquid phase separation and related critical phenomena<sup>18–20</sup>.

In agreement with predictions from mode-coupling theory<sup>9</sup>, there is also evidence for a glass or gel transition at low particle volume fractions and high interparticle attractions. Such a scenario obviously affects the ability to form the high quality crystals required for protein crystallography<sup>15</sup>. Using two apparently quite different model systems, we demonstrate the generality of this emerging description of the effect of a short-range attraction combined with either a hard or soft repulsion on the phase behaviour of a wide range of colloidal suspensions.

We first investigated solutions of the globular protein lysozyme (molecular mass 14.4 kDa, radius  $R_m \approx 1.7$  nm)<sup>17–19</sup>. Using small-angle X-ray (SAXS) and neutron (SANS) scattering, we studied spatial correlations in concentrated solutions at low ionic strength, where the long-range repulsive electrostatic potential is only weakly screened. We then compared these findings with confocal microscopy results using colloid–polymer mixtures, a popular model system with easily tunable interactions. Here we used spherical colloidal particles interacting with a long-range repulsion resulting from a modest charge<sup>21</sup> and a short-range attraction induced by a polymer-mediated ‘depletion effect’<sup>22</sup>.

Figure 1 presents SAXS measurements on a lysozyme concentration series at 5 °C. The normalized scattering intensities  $I(q)/c$ , where  $I(q)$  is the scattering intensity at scattering vector  $q$ , and  $c$  is the protein concentration, show a forward intensity  $I(q \rightarrow 0)/c$  that



**Figure 1** Normalized scattered intensity  $I(q)/c$  and corresponding effective structure factors  $S_{\text{eff}}(q)$ , as obtained by SAXS from lysozyme solutions of different concentrations  $c$ . **a**,  $I(q)/c$  of a dilution series ( $3 \text{ mg ml}^{-1}$  to  $273 \text{ mg ml}^{-1}$ ) at 5 °C. Large open circles represent the  $I(q)/c$  values extrapolated to  $q = 0.1 \text{ nm}^{-1}$  and stars show the projection of the peak maximum onto the  $q$ – $c$  plane. **b**,  $S_{\text{eff}}(q)$  of the concentrated samples in **a**. Concentration ranges from  $36 \text{ mg ml}^{-1}$  (open squares) to  $273 \text{ mg ml}^{-1}$  (filled circles).

decreases with increasing concentration (Fig. 1a). This indicates strongly repulsive interactions, in contrast to efficiently screened lysozyme solutions for which  $I(q \rightarrow 0)/c$  increases because the short-range attractions are dominant<sup>9</sup>. At the same time a pronounced peak at finite  $q$  evolves, indicating strong positional correlations between the proteins.

Further information about these correlations comes from the effective structure factor  $S_{\text{eff}}(q)$ , which is obtained by dividing  $I(q)/c$  by the normalized intensity of a very dilute sample. Qualitatively, a peak in  $S_{\text{eff}}(q)$  at  $q^*$  can be interpreted as a 'Bragg reflection' from planes of particles separated by the mean nearest-neighbour distance  $d$ , with  $d \approx 2\pi/q^*$ . Charged colloids at low ionic strength maximize their average interparticle distance, and so  $d$  depends on the volume fraction  $\phi$  and we find  $q^* \approx 2\pi n^{1/3}$ , where  $n = 3\phi/(4\pi R^3)$  is the number density of particles with radius  $R$ . At the same time the forward scattering becomes suppressed upon increasing  $\phi$ , owing to the decreased osmotic compressibility.

Surprisingly, the peak position  $q_c^*$  of  $S_{\text{eff}}(q)$  is essentially independent of  $\phi$  (Fig. 1b). Moreover, there appears to be an upturn at high  $q$  values. SANS measurements over an extended range of  $q$  values indeed show the existence of a second peak in  $S_{\text{eff}}(q)$  (Fig. 2). The position  $q_m^*$  and height of this second peak changes neither with concentration nor with temperature. The position  $q_c^*$  of the low- $q$  peak is also independent of concentration, in agreement with our SAXS results (Fig. 1b), but it does depend on temperature and shifts to lower  $q$  values with decreasing temperature.

These unexpected observations can be understood if the proteins self-assemble into small clusters with a  $\phi$ -dependent aggregation number  $N_c$ . The driving force for self-assembly into clusters is short-range attraction, which effectively acts as surface tension leading to a decrease in surface energy upon aggregation. On the other hand, cluster growth is limited by the increasing electrostatic energy of the clusters, which counterbalances the gain in surface energy and is due to the small number of residual charges combined with a low ionic strength. A low ionic strength ensures that the Debye length is larger or comparable to the cluster size. This balance

between short-range attraction and weakly screened (and thus long-range) electrostatic repulsion provides a stabilizing mechanism against gelation and determines a finite aggregation number  $N_c$ , as in micelle formation. Equilibrium cluster formation in colloidal suspensions has recently been investigated theoretically and supported by computer simulations<sup>10,11</sup>. Moreover, a theoretical model has previously been proposed for mesophase separation of colloids in organic solvents and an explicit expression for the concentration dependence of  $N_c$  has been derived<sup>12</sup>.

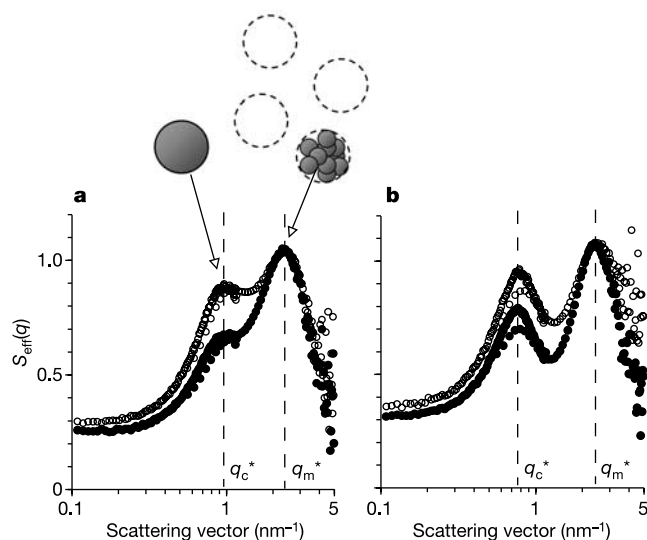
This model allows us to assign the first peak in  $S_{\text{eff}}(q)$  at  $q_c^*$  (Fig. 2a) to cluster-cluster correlations caused by electrostatic interactions between the charged clusters, whereas the second peak at  $q_m^*$  reflects the positional correlations of the monomers within a single cluster. This interpretation is directly supported by numerical simulations<sup>10,11</sup>. The monomer peak at  $q_m^*$  exhibits a concentration and temperature independent value  $q_m^* R_m \approx 3.8$ , indicating a constant packing density within the cluster of approximately 60% by volume<sup>23</sup>. Also the position  $q_c^*$  of the cluster-cluster peak is independent of  $\phi$  at higher concentrations, which implies a constant cluster number density  $n_c$  (because  $q_c^* \propto n_c^{1/3}$ ). This allows us to determine the dependence of  $N_c$  on  $\phi$  according to  $N_c = n_m/n_c \propto \phi/n_c$ , where  $n_m$  is the monomer number density, and the proportionality constant has been determined from a comparison with calculations using the Rogers-Young closure relation<sup>24</sup>. We find  $N_c \propto \phi$ , in agreement with the theoretical prediction<sup>12</sup>. This is not obvious for such small clusters and high volume fractions where we would expect to find considerable contributions to the free energy from both (translational) entropy and from cluster-cluster interactions. However, the entropy contribution that would favour the formation of even smaller clusters may largely be compensated for by the curvature dependence of the interfacial tension. Small clusters have a relatively large number of surface proteins with unfavourable 'broken contacts', and surface tension may thus increase correspondingly.

From  $q_c^*$  and  $\phi$ , absolute values of an average  $N_c$  can be estimated, using the above relationships (Fig. 3a). The effect of temperature can be understood from the subtle balance of opposing forces; decreasing temperature increases the attraction<sup>18</sup> and hence leads to larger clusters and smaller  $n_c$ , which in turn decreases  $q_c^*$ . It is important to point out that the temperature-dependent cluster formation is fully reversible, which strongly supports the idea of equilibrium cluster formation.

Decreasing the temperature shifts the balance in favour of attractive interactions, but a similar effect is also expected following the reduction of electrostatic repulsion by increasing the salt concentration. We did indeed observe this (Fig. 4). The first peak of  $S_{\text{eff}}(q)$  (filled circles) is shifted to a lower  $q$  value when the temperature is decreased (open circles) or the salt concentration is increased (open squares). We estimate  $N_c \approx 3$  at 20 °C without salt, and  $N_c \approx 5$  when either the temperature is decreased to 5 °C or 50 mM NaCl is added. However, temperature and salt are not completely equivalent; salt also screens cluster-cluster interactions, which leads to enhanced forward scattering (Fig. 4). This demonstrates that the subtle balance between short-range attraction and weakly screened Coulomb repulsion allows us to tune the aggregation behaviour and the resulting cluster sizes.

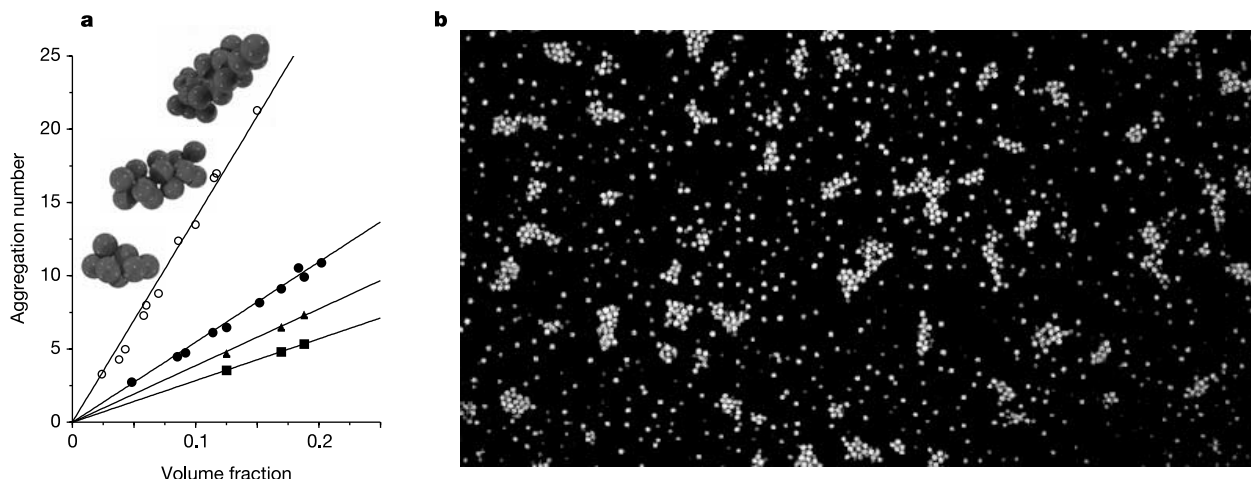
With a further increase in salt concentration, the screening length will eventually become significantly smaller than the cluster size. The electrostatic energy can then no longer stabilize the clusters, leading to irreversible aggregation and precipitation. At 50 mM NaCl, the Debye length is approximately 1.3 nm, and we indeed observe amorphous precipitates at temperatures below 15 °C (Fig. 4, inset).

There are numerous reports of finite size clusters or 'aggregates' of unknown origin in biological and colloidal systems; these include aggregate formation in insulin<sup>16</sup>, the formation of almost monodisperse mixed aggregates of polyelectrolytes (such as DNA) and



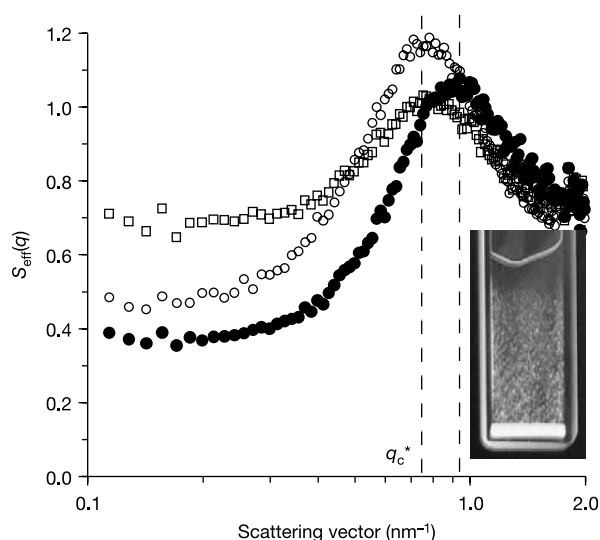
**Figure 2** Effect of concentration and temperature on the effective structure factor  $S_{\text{eff}}(q)$  as obtained by SANS. **a**, 254 mg ml<sup>-1</sup> (filled symbols) and 169 mg ml<sup>-1</sup> (open symbols) lysozyme solutions at 25 °C. **b**, The same samples at 5 °C. The dashed lines highlight that both peak positions are independent of lysozyme concentration. The second peak (corresponding to internal monomer-monomer correlations within the dense particle clusters) changes neither with concentration nor with temperature. The cluster-cluster correlation peak at lower  $q$  is also concentration independent but shows a strong temperature dependence, indicating fewer but larger aggregates at lower temperatures.





**Figure 3** Clusters in protein solutions and colloidal suspensions. **a**, Average aggregation number of clusters ( $N_c$ ) versus volume fraction ( $\phi$ ) in lysozyme samples at 5 °C (filled circles), 15 °C (triangles) and 25 °C (squares), and in colloid–polymer mixtures ( $c_p \approx 3 \text{ mg ml}^{-1}$ , open circles). Values for lysozyme samples were determined using SAXS, and confocal microscopy was used for colloid–polymer mixtures. The insets show

typical clusters plotted using real particle coordinates from confocal microscopy experiments on colloid–polymer mixtures ( $c_p \approx 3 \text{ mg ml}^{-1}$ ,  $\phi = 0.024, 0.086$  and  $0.15$ ). **b**, Confocal microscopy image from a colloid–polymer mixture ( $\phi = 0.086$ ,  $c_p \approx 3 \text{ mg ml}^{-1}$ ).



**Figure 4** Effect of temperature and ionic strength on the effective structure factor  $S_{\text{eff}}(q)$ , obtained by SAXS.  $125 \text{ mg ml}^{-1}$  lysozyme solution without salt (open and filled circles) and with 50 mM NaCl (open squares). The filled circles and open squares represent data obtained at 20 °C, the open circles show data obtained at 5 °C. The dashed lines highlight that increasing the ionic strength or decreasing the temperature both shift the cluster–cluster peak to lower  $q$  values. The inset shows the onset of precipitation in a sample with 50 mM NaCl when the temperature is decreased below 15 °C.

surfactants<sup>14</sup>, and the existence of pre-nucleation clusters in crystallizing protein solutions<sup>17</sup>. We speculate that it is the delicate balance between short-range attractive and electrostatic repulsive forces that leads to this controlled self-assembly.

To explore the general validity of this concept across a broad range of systems and to observe these equilibrium clusters directly, we investigated a colloidal model system using confocal microscopy. We used density-matched, hard-sphere colloidal particles (radius 660 nm) with a modest charge and a short-range (about 2% of the particle diameter) attraction induced by a polymer-mediated ‘depletion effect’.

As the polymer concentration  $c_p$  (and therefore the strength of the attraction) is increased, a transition from an ergodic fluid (in

which individual particles undergo brownian motion) to clustering occurs. Thereafter, over an intermediate range of  $c_p$  values, individual mobile clusters exist (Fig. 3b). At low  $\phi$ , the clusters coexist with single particles, whereas at high  $\phi$  monomers are less frequent. These states of cluster–monomer coexistence seem to be long-lived: clusters were not observed to grow by further aggregation. At even higher  $c_p$ , gelation occurs.

Figure 3a shows the average aggregation number of clusters,  $N_c$ , as a function of colloid volume fraction  $\phi$  at  $c_p \approx 3 \text{ mg ml}^{-1}$ , which corresponds to a contact attraction between neighbours of about  $5k_B T$ , where  $k_B$  is Boltzmann’s constant. As for the protein samples, an  $N_c \propto \phi$  scaling is found, which agrees with the theoretical model and computer simulations<sup>11,12</sup>. Moreover, the absolute values are consistent with the system parameters<sup>12</sup>.  $N_c$  was furthermore observed to rise slightly when  $c_p$  was increased, which is equivalent to deepening the interparticle attraction; this is in agreement with the temperature dependence found for the protein solutions.

Our results from two very different systems unambiguously confirm that the formation of equilibrium clusters in protein solutions is not caused by protein-specific interactions; rather, the combination of a weakly screened, long-range electrostatic repulsion and a short-range attraction leads to the formation of small equilibrium clusters with a concentration-dependent aggregation number  $N_c$ . Our findings demonstrate the general importance of residual, weakly screened charges together with short-range attractions in cluster formation, and suggest a possible means of obtaining tunable cluster formation through protein or colloid self-assembly. This mechanism might also provide an alternative route to nanostructured colloidal gel and glass phases, where the structural elements could be self-assembled colloid clusters<sup>11</sup>. □

## Methods

### Preparation of protein solutions

Hen egg white lysozyme was obtained from Fluka (L7651, three times crystallized, dialysed and lyophilized) and used without further purification. About 40 mg protein was dissolved per millilitre of a 20 mM HEPES buffer in  $D_2O$  (99.9%, Cambridge Isotope Laboratories) at pH 7.8, where the lysozyme carries a net positive charge of about eight electronic charges<sup>19,25</sup>. This stock solution was stirred at room temperature and passed through a 0.22- $\mu\text{m}$  filter to remove any undissolved material. An Amicon ultrafiltration stirring cell with a YM-10 membrane was used to wash the protein solution with buffer (to reduce any ionic impurities left) and to further concentrate it. At the highest lysozyme concentrations, the pH had slightly increased to values between 8.0 and 8.2, corresponding to a charge uncertainty of  $\pm 2\%$  (ref. 25). Control experiments using samples prepared under strict pH control (pH values for all concentrations were  $7.8 \pm 0.1$ ) gave the same

results as the samples with small pH variations, thus indicating that the small variations in the effective net charge have no measurable influence on the cluster formation mechanism (see Supplementary Information). We also checked the temperature dependence of the pH for a concentrated lysozyme solution and found an increase of only 0.2 units when decreasing the temperature from 33 to 5 °C. Again, this results in a negligible charge variation when looking at the titration curve<sup>21</sup>. Lower concentrations were prepared by diluting the stock solution with buffer at pH 7.8. The samples with 50 mM NaCl were obtained by diluting a concentrated protein sample with HEPES buffer containing the appropriate amount of NaCl at pH 7.8. The final concentrations were determined by ultraviolet absorption spectroscopy at 280 nm using a specific absorption coefficient  $E_{1\text{cm}}^{1\%} = 26.4$ ; the highest concentrations were typically between 250 and 350 mg ml<sup>-1</sup>. Using a partial specific volume of 0.74 cm<sup>3</sup> g<sup>-1</sup> for the proteins results in the corresponding protein monomer volume fractions of  $0.185 \leq \phi \leq 0.26$ .

### Preparation of colloid-polymer mixtures

Spherical particles (radius  $R = 660$  nm) with polymethylmethacrylate (PMMA) cores fluorescently labelled with nitrobenzoxadiazole and sterically stabilized by a thin (~10 nm) layer of chemically grafted poly-12-hydroxystearic acid were suspended in an approximately 1:4 mixture of *cis*-decalin and cycloheptyl bromide (CHB) for density matching. CHB leads to a positive charge  $Q$  of the PMMA particles<sup>21</sup> with  $Q \leq 10^3$  electronic charges for the present particles (estimated from  $\phi$  at crystallization). Addition of linear, non-adsorbing polystyrene (Polymer Laboratory, molecular mass 212.4 kDa) induces an effective attraction between the PMMA particles: exclusion of polymer between the surfaces of two nearby particles results in a net osmotic force that pushes them together<sup>22</sup>. The depth and range of this 'depletion' attraction are proportional to the polymer concentration and polymer size, respectively. Polymer size can be estimated by twice the radius of gyration  $r_g$  of a single coil, giving a dimensionless range  $\delta \approx r_g/R$ , here  $\delta \approx 0.02$ .

### Small-angle X-ray scattering measurements

SAXS experiments were carried out with a pinhole camera (NanoSTAR, Bruker AXS) equipped with a sealed tube (Cu K $\alpha$ ), a thermostatically regulated sample chamber and a two-dimensional gas detector. The  $q$  range is 0.1–2 nm<sup>-1</sup>.

### Small-angle neutron scattering measurements

SANS experiments were performed at the SANS I facility at the Swiss neutron source SINQ at the Paul Scherrer Institut, Switzerland. We used 1-mm and 2-mm Hellma quartz cells and a thermostatically regulated sample holder. Combinations of different wavelengths (5 and 8 Å), sample-to-detector distances (1.6–18 m) and collimation lengths (4.5–18 m) were used to cover a  $q$  range of 0.1–7 nm<sup>-1</sup>.

### Confocal microscopy

Imaging was carried out in the Collaborative Optical Spectroscopy, Micromanipulation and Imaging Centre (COSMIC). A small amount of sample was sandwiched between a cover slip and a microscope slide. The sample thickness was fixed at about 0.3 mm by spacers. A Nikon TE-300 with a Biorad Radiance 2100MP scanning head was used; fluorescence was excited at 488 nm and observed at 525 nm.

Received 1 September; accepted 6 October 2004; doi:10.1038/nature03109.

- Dawson, K. A. The glass paradigm for colloidal glasses, gels, and other arrested states driven by attractive interactions. *Curr. Opin. Colloid Interf. Sci.* **7**, 218–227 (2002).
- Trappe, V., Prasad, V., Cipelletti, L., Segre, P. N. & Weitz, D. A. Jamming phase diagram for attractive particles. *Nature* **411**, 772–775 (2001).
- Sciortino, F. Disordered materials: one liquid, two glasses. *Nature Mater.* **1**, 145–146 (2002).
- Pham, K. N. *et al.* Multiple glassy states in a simple model system. *Science* **296**, 104–106 (2002).
- Eckert, T. & Bartsch, E. Re-entrant glass transition in a colloid-polymer mixture with depletion attractions. *Phys. Rev. Lett.* **89**, 125701–125704 (2002).
- Weeks, E. R., Crocker, J. C., Levitt, A. C., Schofield, A. & Weitz, D. A. Three-dimensional direct imaging of structural relaxation near the colloidal glass transition. *Science* **287**, 627–631 (2000).
- Foffi, G. *et al.* Phase equilibria and glass transition in colloidal systems with short-ranged attractive interactions: application to protein crystallization. *Phys. Rev. E* **65**, 031407–031417 (2002).
- Bergenholtz, J., Poon, W. C. K. & Fuchs, M. Gelation in model colloid-polymer mixtures. *Langmuir* **19**, 4493–4503 (2003).
- Kulkarni, A. M., Dixit, N. M. & Zukoski, C. F. Ergodic and non-ergodic phase transitions in globular protein suspensions. *Faraday Discuss.* **123**, 37–50 (2003).
- Puertas, A. M., Fuchs, M. & Cates, M. E. Dynamical heterogeneities close to a colloidal gel. *J. Chem. Phys.* **121**, 2813–2822 (2004).
- Sciortino, F., Mossa, S., Zaccarelli, E. & Tartaglia, P. Equilibrium cluster phases and low-density arrested disordered states: The role of short-range attraction and long-range repulsion. *Phys. Rev. Lett.* **93**, 055701 (2004).
- Groenewold, J. & Kegel, W. K. Anomalous large equilibrium clusters of colloids. *J. Phys. Chem. B* **105**, 11702–11709 (2001).
- Segré, P. N., Prasad, V., Schofield, A. B. & Weitz, D. A. Glasslike kinetic arrest at colloidal-gelation transition. *Phys. Rev. Lett.* **86**, 6042–6045 (2001).
- Guillot, S., Delsanti, M., Désert, S. & Langevin, D. Surfactant-induced collapse of polymer chains and monodisperse growth of aggregates near the precipitation boundary in carboxymethylcellulose-DTAB aqueous solutions. *Langmuir* **19**, 230–237 (2003).
- Muschol, M. & Rosenberger, F. Liquid-liquid phase separation in supersaturated lysozyme solutions and associated precipitate formation/crystallization. *J. Chem. Phys.* **107**, 1953–1962 (1997).
- Pedersen, J. S., Hansen, S. & Bauer, R. The aggregation behavior of zinc-free insulin studied by small-angle neutron scattering. *Eur. Biophys. J.* **22**, 379–389 (1994).
- Piazza, R. Interactions and phase transitions in protein solutions. *Curr. Opin. Colloid Interf. Sci.* **5**, 38–43 (2000).

- Malfois, M., Bonnete, F., Belloni, L. & Tardieu, A. A model of attractive interactions to account for fluid-fluid phase separation of protein solutions. *J. Chem. Phys.* **105**, 3290–3300 (1996).
- Broide, M. L., Tomic, T. M. & Saxowsky, M. D. Using phase transitions to investigate the effect of salts on protein interactions. *Phys. Rev. E* **53**, 6325–6335 (1996).
- Schurtenberger, P., Chamberlin, R. A., Thurston, G. M., Thomson, J. A. & Benedek, G. B. Observation of critical phenomena in a protein-water solution. *Phys. Rev. Lett.* **63**, 2064–2067 (1989).
- Yethiraj, A. & Van Blaaderen, A. A colloidal model system with an interaction tunable from hard sphere to soft and dipolar. *Nature* **421**, 513–517 (2003).
- Poon, W. C. K. The physics of a model colloid-polymer mixture. *J. Phys. Condens. Matter* **14**, R859–R880 (2002).
- Pham, K. N., Egelhaaf, S. U., Pusey, P. N. & Poon, W. C. K. Glasses in hard spheres with short-range attraction. *Phys. Rev. E* **69**, 11503–11516 (2004).
- Rojas, L., Urban, C., Schurtenberger, P., Gisler, T. & Grünberg, H. H. Reappearance of structure in charge-stabilized suspensions. *Europhys. Lett.* **60**, 802–808 (2002).
- Tanford, C. & Roxy, R. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry* **11**, 2192–2198 (1972).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank the Swiss spallation source at the Paul Scherrer Institut (PSI) in Villigen, Switzerland, for the neutron beam time and we acknowledge the help of our local contacts J. Kohlbrecher and S. van Petegem. We thank J. Groenewold, W. Kegel, F. Sciortino, K. Kroy and M. Cates for discussions. We thank A. Schofield for preparing the fluorescent PMMA particles. This work was supported by the Swiss National Science Foundation, the UK Engineering and Physical Sciences Research Council, the Scottish Higher Education Funding Council, and the Marie Curie Network on Dynamical Arrest of Soft Matter and Colloids. A.S. and P.S. conceived and performed the protein experiments; F.C. prepared the pH stabilized protein samples for the control experiments; H.S., W.C.K.P. and S.U.E. carried out and analysed the experiments with the colloid-polymer samples.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to P.S. ([peter.schurtenberger@unifr.ch](mailto:peter.schurtenberger@unifr.ch)).

## A humid climate state during the Palaeocene/Eocene thermal maximum

Gabriel J. Bowen<sup>1\*</sup>, David J. Beerling<sup>2</sup>, Paul L. Koch<sup>1</sup>, James C. Zachos<sup>1</sup> & Thomas Quattlebaum<sup>1</sup>

<sup>1</sup>Earth Sciences Department, University of California, Santa Cruz, California 95064, USA

<sup>2</sup>Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

\* Present address: Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA

An abrupt climate warming of 5 to 10 °C during the Palaeocene/Eocene boundary thermal maximum (PETM) 55 Myr ago is linked to the catastrophic release of ~1,050–2,100 Gt of carbon from sea-floor methane hydrate reservoirs<sup>1</sup>. Although atmospheric methane, and the carbon dioxide derived from its oxidation, probably contributed to PETM warming, neither the magnitude nor the timing of the climate change is consistent with direct greenhouse forcing by the carbon derived from methane hydrate. Here we demonstrate significant differences between marine<sup>2,3</sup> and terrestrial<sup>4–6</sup> carbon isotope records spanning the PETM. We use models of key carbon cycle processes<sup>7–9</sup> to identify the cause of these differences. Our results provide evidence for a previously unrecognized discrete shift in the state of the climate system during the PETM, characterized by large increases in mid-latitude tropospheric humidity and enhanced cycling of carbon through terrestrial ecosystems. A more humid atmosphere helps to explain PETM temperatures, but the ultimate mechanisms underlying the shift remain unknown.

Global warming during the PETM is associated with a major negative carbon isotope ( $\delta^{13}\text{C}$ , see Fig. 1 legend) excursion (CIE), which has been invoked as evidence for the release of  $^{13}\text{C}$ -poor carbon from methane hydrate reservoirs into the ocean and atmosphere<sup>1</sup>. The signature of this carbon release is a rapid and synchronous decrease in the  $\delta^{13}\text{C}$  of carbon in terrestrial<sup>4–6,10</sup> and marine<sup>2,3</sup> rocks (Fig. 1). The  $\delta^{13}\text{C}$  values stabilize approximately 35 kyr after the beginning of the event, indicating the cessation of methane release. Throughout the following  $\sim 45$  kyr,  $\delta^{13}\text{C}$  values remain relatively low and then begin an exponential recovery lasting for some  $\sim 50$  kyr. Temperature, in contrast, rises steadily from the beginning of the event to a peak  $\sim 60$  kyr later, and gradually declines to pre-PETM levels over the next 70 kyr. The release of  $\text{CH}_4$  into the atmosphere during the PETM could have played an important role in warming the climate<sup>11,12</sup>, but because of its short residence time  $\text{CH}_4$  can not explain the continued rise in temperatures following the termination of methane release. Atmospheric methane oxidizes to  $\text{CO}_2$ , which has a much longer residence time, but this  $\text{CO}_2$  would have increased levels in the PETM atmosphere by only 70 to 160 p.p.m.v. (refs 1, 11), raising global temperature by less than  $1^\circ\text{C}$  (ref. 12). These observations suggest that during the PETM the climate responded nonlinearly to changes in radiative forcing resulting from the addition of hydrate-derived carbon to the atmosphere.

Recently developed high-resolution stratigraphic records<sup>3–6</sup> and timescales<sup>13</sup> allow comparison of terrestrial and marine  $\delta^{13}\text{C}$  records at the global scale. Important isotopic records are available for many sites, but we focus here on well-sampled records derived from characterized substrates that document the full shift in  $\delta^{13}\text{C}$  values from a clearly defined, pre-PETM baseline through the CIE. These time series show that during a discrete, 60-kyr interval that includes peak PETM warming, the CIE in palaeosol carbonates from Wyoming (USA), Spain and China is consistently amplified by  $\sim 3\text{‰}$  relative to that in marine records (Fig. 1d). Fossil soil organic matter (SOM) from terrestrial rocks in northern Wyoming<sup>10</sup> also documents an excursion  $\sim 1\text{‰}$  larger than the marine CIE. This feature is common to the terrestrial  $\delta^{13}\text{C}$  record on three continents, which strongly suggests that it is not a diagenetic artefact but reflects a discrete change in the partitioning of  $^{13}\text{C}$

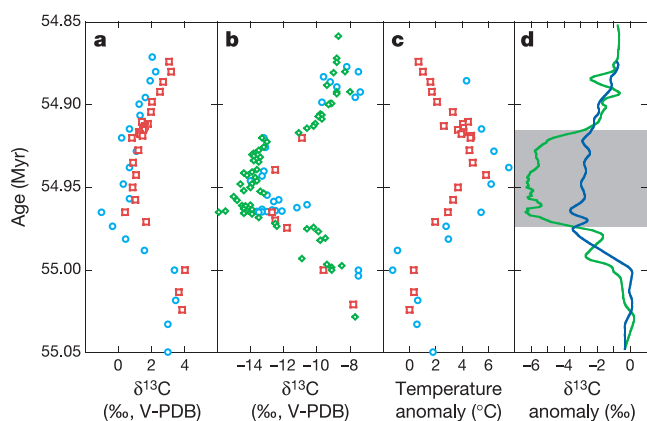
among the ocean, atmosphere, terrestrial biosphere and soils coinciding with PETM nonlinear climate system changes.

Here we integrate data and models to identify the carbon cycle processes responsible for amplifying the terrestrial CIE signal. Our analyses show that several processes could have made minor contributions to terrestrial CIE amplification. These include: (1) temperature effects on carbon isotope fractionation between  $\text{CO}_2$  gas, dissolved inorganic carbon and  $\text{CaCO}_3$  in the surface ocean and in soils; (2) changes in surface ocean carbonate ion concentration ( $[\text{CO}_3^{2-}]$ ); and (3) changes in soil productivity and organic matter turnover rates.

Estimates of PETM warming from terrestrial<sup>14</sup> and surface ocean<sup>2,3</sup> records both fall between  $5$  and  $10^\circ\text{C}$ , implying approximately equal temperature-driven decreases in palaeosol and foraminiferal carbonate  $\delta^{13}\text{C}$  ( $\delta^{13}\text{C}_{\text{PC}}$  and  $\delta^{13}\text{C}_{\text{FC}}$  respectively) of  $0.55$ – $1.1\text{‰}$  (ref. 15). Thus, the temperature effect exaggerates the magnitude of the CIE in these carbonate records relative to the shift in the other exogenic reservoirs, but does not contribute substantially to the offset between marine and terrestrial carbonate records.

Carbon release at the PETM would have affected ocean carbonate chemistry, causing a decrease in ocean pH and  $[\text{CO}_3^{2-}]$  and an increase in  $\delta^{13}\text{C}_{\text{FC}}$  (ref. 8). Mass balance calculations (see Supplementary Information) indicate that carbon addition to the ocean/atmosphere during the PETM could cause a  $0.2$ – $0.6\text{‰}$  increase in  $\delta^{13}\text{C}_{\text{FC}}$ , thereby decreasing the CIE amplitude in marine records. If the mass of carbon released during the PETM has been severely underestimated<sup>16,17</sup>, the  $[\text{CO}_3^{2-}]$  effect may have been even larger. However, the  $\delta^{13}\text{C}$  record of deep ocean foraminifera constrains the  $[\text{CO}_3^{2-}]$  effect. Deepwater  $[\text{CO}_3^{2-}]$  is less sensitive to carbon addition than surface water  $[\text{CO}_3^{2-}]$  because of its greater overall carbon concentration and lower pH, and the  $[\text{CO}_3^{2-}]$  effect should increase the  $\delta^{13}\text{C}$  of surface ocean foraminifera more than that of deepwater forms. In fact, the CIE amplitude for surface-dwelling foraminifera during the PETM is larger than that for deep-dwellers. This observation has been attributed to changes in ocean circulation or productivity<sup>2</sup> but also suggests that  $[\text{CO}_3^{2-}]$  change did not have an extreme effect on marine CIE records.

Soil processes determining the  $\delta^{13}\text{C}$  offset between plants, SOM and soil  $\text{CO}_2$  probably varied during the PETM. Experimental work has shown that warming and  $\text{CO}_2$ -fertilization can lead to increased rates of carbon input to soils (as root and leaf litter)<sup>18</sup> as well as increased turnover rates for SOM<sup>19</sup>. The  $\delta^{13}\text{C}$  of soil carbonate is largely determined by that of soil  $\text{CO}_2$  (ref. 20) and is sensitive to the rates of carbon input and SOM turnover, because they affect both the  $\delta^{13}\text{C}$  of SOM ( $\delta^{13}\text{C}_{\text{SOM}}$ , from which most soil  $\text{CO}_2$  is derived), and the mixing ratio of  $^{13}\text{C}$ -enriched atmospheric and  $^{13}\text{C}$ -depleted respired  $\text{CO}_2$  within soil pores. We used a model of SOM and soil  $\text{CO}_2$  dynamics<sup>9</sup> to examine the effects of changes in soil carbon processes on  $\delta^{13}\text{C}_{\text{PC}}$  and  $\delta^{13}\text{C}_{\text{SOM}}$  during the PETM (see Methods). Our simulations show that  $\delta^{13}\text{C}_{\text{PC}}$  decreases in response to increased rates of carbon input to soils, and increases in response to increased turnover rates of SOM, largely as the result of changes in soil  $\text{CO}_2$  concentration as more or less vegetation-derived  $\text{CO}_2$  is concentrated in soil pores (Fig. 2a, b). In contrast,  $\delta^{13}\text{C}_{\text{SOM}}$  is insensitive to the rate of carbon input to soil and increases in response to increased SOM turnover. These processes also cause substantial changes in the organic carbon content of soils (Fig. 2c), allowing us to constrain our simulations with measurements of the organic carbon concentrations of palaeosols from PETM and non-PETM strata (see Supplementary Information). Our results show no significant increase in carbon storage in PETM palaeosols, suggesting that increased organic inputs to PETM soils were balanced by increased SOM turnover rates. If ‘reasonable’ increases in ecosystem productivity during the PETM (that is, up to a doubling of SOM production rates) are combined with increased turnover so that soil carbon storage remains unchanged, our model

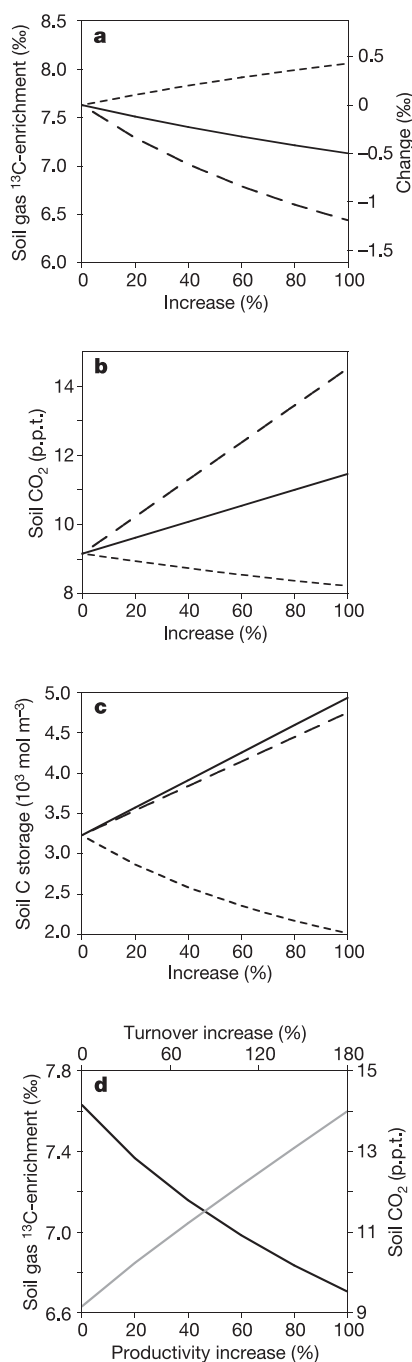


**Figure 1** Marine and terrestrial records of the PETM, correlated to an age model for ODP site 690 (ref. 13, see Supplementary Information). **a**, Marine  $\delta^{13}\text{C}$  records derived from the surface-dwelling genus *Acarinina* at ODP sites 690 (ref. 2, Southern Ocean, blue circles) and 1209 (ref. 3, subtropical Pacific Ocean, red squares). **b**, Palaeosol carbonate  $\delta^{13}\text{C}$  records from northern Spain<sup>6</sup> (blue circles), Hunan, China<sup>5</sup> (red squares) and Wyoming, USA<sup>4,30</sup> (green diamonds). **c**, Temperature anomalies for sites 690 and 1209 calculated from monospecific  $\delta^{18}\text{O}$  and  $\text{Mg}/\text{Ca}$  records, respectively (symbols as in **a**). **d**, Normalized composite carbon isotope curves for palaeosol carbonates (green) and planktonic foraminiferal carbonate (dark blue). Interval of terrestrial CIE amplification is shown in grey.  $\delta^{13}\text{C} = \{[(^{13}\text{C}/^{12}\text{C})_{\text{sample}} / (^{13}\text{C}/^{12}\text{C})_{\text{standard}}] - 1\} \times 1,000$ .



indicates that soil carbon dynamics can account for up to 0.9‰ of the terrestrial CIE amplification recorded by palaeosol carbonate (Fig. 2d), but at the same time cause  $\delta^{13}\text{C}_{\text{SOM}}$  to increase by 0.6–1.0‰ (not shown).

In sum, the above effects account for half of the 3‰ CIE amplification in palaeosol carbonate, but leave a 1.7–2.1‰ difference between the modelled and observed SOM amplitudes.



**Figure 2** Results from soil carbon model runs. **a–c**, Effect of changing productivity above ground (solid black line) or below ground (long dashes) and of changing SOM decomposition rate (short dashes) on soil gas  $\delta^{13}\text{C}$  (**a**), soil gas  $\text{CO}_2$  concentration (**b**) and soil C storage (**c**). **d**, Effect of simultaneous changes in productivity and decomposition rate that result in no net increase in soil carbon storage. Soil gas  $^{13}\text{C}$ -enrichment (black line) is the difference between soil  $\delta^{13}\text{C}_{\text{CO}_2}$  at 1 m depth and vegetation  $\delta^{13}\text{C}$ . Soil  $\text{CO}_2$  concentration (grey line) is given for 1 m depth, and organic carbon storage is integrated over the whole soil.

This implies that terrestrial plants increased their photosynthetic  $^{13}\text{C}$ -discrimination by 1.5–2.1‰ during the PETM. We examined effects of changing water stress, the primary determinant of photosynthetic  $^{13}\text{C}$ -discrimination, on the  $\delta^{13}\text{C}$  values of PETM plants (see Methods). Temperature, relative humidity and soil water availability set the level of water stress. Higher PETM temperatures would have increased the transpiration demand on plants, decreasing  $^{13}\text{C}$ -discrimination by 2–3‰ (Fig. 3). According to our simulations, it is only through substantial increases in relative humidity and soil moisture that plant  $^{13}\text{C}$ -discrimination could have decreased despite climatic warming during PETM. We calculate that a minimum 20% increase in soil moisture and relative humidity would have been required during the PETM to account for the 1.5–2.1‰ increase in plant  $^{13}\text{C}$ -discrimination (Fig. 3). About 85% of this change in vegetation  $\delta^{13}\text{C}$  would be transferred to soil  $\text{CO}_2$  and carbonate; a loss of 15% of the signal occurs owing to dilution by atmosphere-derived  $\text{CO}_2$  within soils.

Taken together, our analysis of marine and terrestrial carbon cycle processes provides a coherent explanation for PETM CIE amplification in terrestrial SOM and palaeosol carbonate (Table 1). Notably, the solution requires a 20–25% increase in soil and atmospheric moisture throughout the northern mid-latitudes and a near doubling in the rate of carbon cycling through terrestrial ecosystems. These changes are consistent with clay mineral records that suggest enhanced continental weathering across these regions during the PETM<sup>21</sup>. Given the protracted duration and stable magnitude of the terrestrial CIE amplification, these changes do not fit a model of linear response to methane release, but rather seem to represent a discrete, transient switch in climate state. The cause and effect relationships between methane hydrate destabilization and the PETM climate state switch, as well as the climate system changes underlying the PETM wet climate state, are currently unknown. Although atmospheric moisture is a powerful greenhouse gas and may have contributed to PETM warming, it is an internal component of the climate system. Persistent, elevated relative humidity during the PETM must represent a feedback responding to some other change in the climate system. This change remains to be identified, but potential candidates include a change in ocean circulation and heat transport<sup>12</sup>, higher levels of atmospheric  $\text{CO}_2$  due to changes in ocean circulation and chemistry<sup>2</sup>, or higher  $\text{CH}_4$  concentrations sustained by elevated fluxes from wetlands.

Recognition of a discrete climate state shift during the PETM has important implications for understanding the evolution of greenhouse climate at the Palaeocene/Eocene boundary and the potential evolution of future climate. Climate system changes associated with the PETM wet climate state help resolve the discrepancy between the observed temperature changes and forcing mechanisms suggested thus far. Both the buildup of tropospheric water vapour and lack of carbon sequestration by soils would have amplified and helped to sustain PETM warmth. In contrast, increases in plant productivity, soil  $\text{CO}_2$  and soil moisture associated with the PETM wet climate state would have increased silicate weathering and the delivery of nutrients and alkalinity to the oceans, increasing the burial of

**Table 1** Proposed model for PETM terrestrial CIE amplification

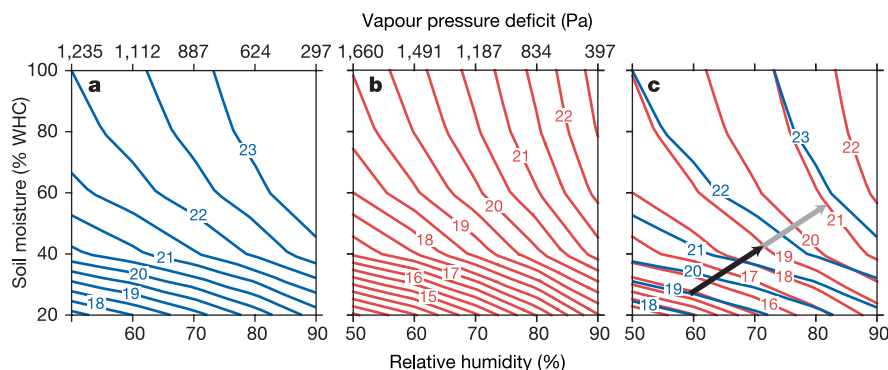
|                                          | $\Delta(\delta^{13}\text{C}_{\text{V-FC}})^*$ | $\Delta(\delta^{13}\text{C}_{\text{SOM-FC}})^*$ | $\Delta(\delta^{13}\text{C}_{\text{PC-FC}})^*$ |
|------------------------------------------|-----------------------------------------------|-------------------------------------------------|------------------------------------------------|
| 5 °C warming†                            | –0.55‰                                        | –0.55‰                                          | 0‰                                             |
| $[\text{CO}_2]$ effect                   | 0.5‰                                          | 0.5‰                                            | 0.5‰                                           |
| Balanced $1.8 \times$ soil productivity‡ | 0‰                                            | –1.0 to –0.6‰                                   | 0.8‰                                           |
| +20% RH and soil moisture                | 2.0‰                                          | 2.0‰                                            | 1.7‰                                           |
| Total                                    | 1.95‰                                         | 0.95 to 1.35‰                                   | 3.0‰                                           |
| Observed                                 | n.a.                                          | ~1.0‰                                           | ~3.0‰                                          |

RH, relative humidity; n.a., not available

\*Change in  $\delta^{13}\text{C}$  offset between terrestrial vegetation (V), soil organic matter (SOM) or palaeosol carbonate (PC) and surface ocean foraminiferal carbonate (FC).

†Equal warming assumed at marine and terrestrial sites.

‡Increased litter input to soils balanced by increased SOM turnover rate.



**Figure 3** Photosynthetic  $^{13}\text{C}$ -discrimination by tropical evergreen  $\text{C}_3$  plants. Isotopic discrimination (labelled colour contours, in ‰) is given for temperatures of 20 °C (pre-PETM, **a**) and 25 °C (PETM, **b**) and a range of atmospheric and soil moisture conditions. **c**, Discrimination shown for both temperatures superimposed on common axes for relative humidity and soil moisture (colours as in **a**, **b**). At both mean annual temperatures, photosynthetic  $^{13}\text{C}$ -discrimination increases (plant  $\delta^{13}\text{C}$  decreases) in

response to increased relative humidity or soil moisture (as a percentage of soil water holding capacity, WHC). Arrows in **c** show increases in relative humidity (black) and soil moisture (grey) required to hold plant  $\delta^{13}\text{C}$  constant at the estimated pre-PETM value (19‰, see Supplementary Information) and to account for the observed 1.5–2.1‰ decrease in plant  $\delta^{13}\text{C}$  during the PETM.

carbon in marine carbonate rocks<sup>13</sup>. This negative feedback probably contributed to climate recovery following the PETM, but temperature stabilization and recovery lagged the climate state switch by ~30–40 kyr. Given the current exponential increase in atmospheric  $\text{CO}_2$  and other greenhouse gases, determining the complete sequence of events associated with the PETM climate state change and assessing the generality of the state change through study of similar episodes of greenhouse warming in the Earth's history<sup>22,23</sup> seems critical. □

## Methods

### Soil carbon model

The steady-state model for soil organic carbon and  $\text{CO}_2$  was described in ref. 9. It includes input functions for organic carbon above and below ground and for root-respired  $\text{CO}_2$ , along with vertical transport, microbially mediated transformation of organic carbon among three discrete organic carbon pools, heterotrophic respiration, and biological and physical  $^{13}\text{C}$ -fractionating processes associated with microbial respiration and diffusion of  $\text{CO}_2$ . For initial conditions, we used model parameter values fitted to data from a carbonate-bearing chernozem<sup>9</sup> and plant  $^{13}\text{C}$ -discrimination ( $\Delta$ ) = 19‰. In the initial conditions, the model soil stores 3,229  $\text{mol C m}^{-2}$  and at 1 m depth has a  $\text{CO}_2$  concentration of 9.16 parts per thousand (p.p.t.) with soil gas  $\delta^{13}\text{C}$  enriched by 7.6‰ relative to vegetation, bulk SOM enriched by 2.4‰ relative to vegetation, and stable<sup>9</sup> SOM enriched by 2.9‰ relative to vegetation. We tested the sensitivity of soil  $\text{CO}_2$  concentration and  $\delta^{13}\text{C}$  values, and also SOM storage and  $\delta^{13}\text{C}$  values, to changes in model parameters by incrementally changing them from their initial values, either individually or together, and re-integrating the model. The tests reported here were performed using ranges of  $\Delta$  values (19–23‰), plant productivity rates above ground (30 to 60  $\text{mol C m}^{-2} \text{ yr}^{-1}$ ) and below ground (20 to 40  $\text{mol C m}^{-2} \text{ yr}^{-1}$ ), and soil carbon turnover rate (0.2 to 0.4, 0.01 to 0.02, and 0.001 to 0.002  $\text{yr}^{-1}$  for carbon cycling at 'fast', 'slow' and 'stable' rates, respectively). These tests assume that root respiration varies proportionately with productivity below ground. We also tested the effects of decreased microbial assimilation efficiency<sup>9</sup> (0.4 to 0.2  $\text{mol C}$  assimilated per  $\text{mol C}$  consumed for fast and slow cycling carbon, and 0.2 to 0.1 for stable carbon levels); the results were very similar to those obtained with increased turnover rate and are not shown here. All other model parameters were held constant, and simulations were run at an atmospheric  $\text{CO}_2$  concentration of 1,500 p.p.m. Results reported here are for depth-integrated SOM storage and for soil gas and bulk and stable SOM at 1 m below the soil surface.

### Modelling plant carbon isotope fractionation

Discrimination against  $^{13}\text{C}$  ( $\Delta$ ) by leaves was calculated using the well-validated model<sup>7</sup> linking  $\Delta$  to leaf gas exchange, which is given as:  $\Delta = a + (b - a) \times c_i/c_a$ , where  $a$  is fractionation associated with diffusion (4.4‰),  $b$  is fractionation associated with the enzyme Rubisco (27‰), and  $c_i$  and  $c_a$  are the intercellular and atmospheric  $\text{CO}_2$  partial pressures respectively. Experimental and geological evidence indicates that terrestrial plant  $\delta^{13}\text{C}$  values are insensitive to changes in  $\text{CO}_2$  (ref. 24), but that water stress is the primary factor determining the magnitude of terrestrial plant photosynthetic  $^{13}\text{C}$ -discrimination<sup>7</sup>. We calculated equilibrium  $c_i/c_a$  ratios at  $c_a = 1,500$  p.p.m.v. over a range of humidity, temperature and soil moisture content values using a mechanistic model of photosynthetic carbon uptake<sup>25</sup> coupled to a model of stomatal behaviour<sup>26</sup>. The coupled model incorporated the effects of leaf-to-air difference in the molar concentration of water vapour on stomatal conductance and accounted for soil moisture based on the reductions in stomatal conductance that occur with soil drying<sup>27</sup>. All simulations used maximum

rates of carboxylation activity ( $94.1 \mu\text{mol m}^{-2} \text{ s}^{-1}$ ) and photosynthetic electron transport ( $183.1 \mu\text{mol m}^{-2} \text{ s}^{-1}$ ), characteristic of the deciduous and evergreen tropical forests<sup>28</sup> that predominated at the continental PETM sites investigated<sup>29</sup>.

Received 25 March; accepted 12 October 2004; doi:10.1038/nature03115.

- Dickens, G. R., Castillo, M. M. & Walker, J. C. G. A blast of gas in the latest Paleocene; simulating first-order effects of massive dissociation of oceanic methane hydrate. *Geology* **25**, 259–262 (1997).
- Kennett, J. P. & Stott, L. D. Abrupt deep-sea warming, palaeoceanographic changes and benthic extinctions at the end of the Paleocene. *Nature* **353**, 225–229 (1991).
- Zachos, J. C. *et al.* A transient rise in tropical sea surface temperature during the Paleocene-Eocene thermal maximum. *Science* **302**, 1551–1554 (2003).
- Bowen, G. J. *et al.* in *Paleocene-Eocene Stratigraphy and Biotic Change in the Bighorn and Clarks Fork Basins, Wyoming* (ed. Gingerich, P. D.) 73–88 (Univ. of Michigan Museum of Paleontology, Ann Arbor, Michigan, 2001).
- Bowen, G. J. *et al.* Mammalian dispersal at the Paleocene/Eocene boundary. *Science* **295**, 2062–2065 (2002).
- Schmitz, B. & Pujalte, V. Sea-level, humidity, and land-erosion records across the initial Eocene thermal maximum from a continental-marine transect in northern Spain. *Geology* **31**, 689–692 (2003).
- Farquhar, G. D., Ehleringer, J. R. & Hubrick, K. T. Carbon isotope discrimination and photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **40**, 503–537 (1989).
- Spero, H. J., Bijma, J., Lea, D. W. & Bemis, B. E. Effect of seawater carbonate concentration on foraminiferal carbon and oxygen isotopes. *Nature* **390**, 497–500 (1997).
- Bowen, G. J. & Beerling, D. J. An integrated model for soil organic carbon and  $\text{CO}_2$ : implications for paleosol carbonate  $p\text{CO}_2$  paleobarometry. *Glob. Biogeochem. Cycles* **18**, doi:10.1029/2003GB002117 (2004).
- Magioncalda, R., Dupuis, C., Smith, T., Steurbaut, E. & Gingerich, P. D. Paleocene-Eocene carbon isotope excursion in organic carbon and pedogenic carbonate: Direct comparison in a continental stratigraphic section. *Geology* **32**, 553–556 (2004).
- Schmidt, G. A. & Shindell, D. T. Atmospheric composition, radiative forcing, and climate change as a consequence of a massive methane release from gas hydrates. *Paleoceanography* **18**, doi:10.1029/2002PA000757 (2003).
- Renssen, H., Beets, C. J., Fichet, T., Goosse, H. & Kroon, D. Modeling the climate response to a massive methane release from gas hydrates. *Paleoceanography* **19**, doi:10.1029/2003PA000968 (2004).
- Farley, K. A. & Eltgroth, S. F. An alternative age model for the Paleocene-Eocene thermal maximum using extraterrestrial He-3. *Earth Planet. Sci. Lett.* **208**, 135–148 (2003).
- Fricke, H. C. & Wing, S. L. Oxygen isotope and paleobotanical estimates of temperature and  $\delta^{18}\text{O}$ -latitude gradients over North America during the Early Eocene. *Am. J. Sci.* **304**, 612–635 (2004).
- Friedman, I. & O'Neill, J. R. in *Compilation of Stable Isotope Fractionation Factors of Geochemical Interest* (ed. Fleischer, M.) 1–12 (US Geological Survey, Reston, Virginia, 1977).
- Kurtz, A. C., Kump, L. R., Arthur, M. A., Zachos, J. C. & Paytan, A. Early Cenozoic decoupling of the global carbon and sulfur cycles. *Paleoceanography* **18**, doi:10.1029/2003PA000908 (2003).
- Svensen, H. *et al.* Release of methane from a volcanic basin as a mechanism for initial Eocene global warming. *Nature* **429**, 542–545 (2004).
- Schäfer, K. V. R. *et al.* Exposure to an enriched  $\text{CO}_2$  atmosphere alters carbon assimilation and allocation in a pine forest ecosystem. *Glob. Change Biol.* **9**, 1378–1400 (2003).
- Macdonald, N. W., Zak, D. R. & Pregitzer, K. S. Temperature effects on kinetics of microbial respiration and net nitrogen and sulfur mineralization. *Soil Sci. Soc. Am. J.* **59**, 233–240 (1995).
- Cerling, T. E. The stable isotopic composition of modern soil carbonate and its relationship to climate. *Earth Planet. Sci. Lett.* **71**, 229–240 (1984).
- Bolle, M. P. & Adatte, T. Palaeocene-early Eocene climatic evolution in the Tethyan realm; clay mineral evidence. *Clay Miner.* **36**, 249–261 (2001).
- Gröcke, D. R., Hesselbo, S. P. & Jenkyns, H. C. Carbon-isotope composition of Lower Cretaceous fossil wood: ocean-atmosphere chemistry and relation to sea-level change. *Geology* **27**, 155–158 (1999).
- Hesselbo, S. P. *et al.* Massive dissociation of gas hydrate during a Jurassic oceanic anoxic event. *Nature* **406**, 392–395 (2000).
- Tu, T. T. N., Kürschner, W. M., Schouten, S. & Van Bergen, P. F. Leaf carbon isotope composition of

- fossil and extant oaks grown under differing atmospheric CO<sub>2</sub> levels. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **212**, 199–213 (2004).
25. Farquhar, G. D., von Caemmerer, S. & Berry, J. A. A biochemical model of photosynthetic CO<sub>2</sub> assimilation in leaves of C3 species. *Planta* **149**, 78–90 (1980).
  26. Leuning, R. A critical appraisal of a combined stomatal-photosynthesis model for C3 plants. *Plant Cell Environ.* **18**, 339–355 (1995).
  27. Granier, A., Biron, P., Bréda, N., Pontallier, J. Y. & Saugier, B. Transpiration of trees and forest stands: short-term and long-term monitoring using sapflow methods. *Glob. Change Biol.* **2**, 265–274 (1996).
  28. Beerling, D. J. & Quick, W. P. A new technique for estimating rates of carboxylation and electron transport in leaves of C-3 plants for use in dynamic global vegetation models. *Glob. Change Biol.* **1**, 289–294 (1995).
  29. Sewall, J. O., Sloan, L. C., Huber, M. & Wing, S. Climate sensitivity to changes in land surface characteristics. *Glob. Planet. Change* **26**, 445–465 (2000).
  30. Koch, P. L., Zachos, J. C. & Dettman, D. L. Stable isotope stratigraphy and paleoclimatology of the Paleogene Bighorn Basin (Wyoming, USA). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **115**, 61–89 (1995).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank W. Cheng, G. Dickens, D. Schrag, L. Sloan and F. I. Woodward for comments. Funding was provided by a National Science Foundation Biocomplexity grant. G.J.B. was supported by the National Science Foundation Graduate Research Fellowship Program, and D.J.B. gratefully acknowledges funding from the Royal Society and the Leverhulme Trust.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to G.J.B. ([gbowen@biology.utah.edu](mailto:gbowen@biology.utah.edu)).

## Indirect reciprocity can stabilize cooperation without the second-order free rider problem

Karthik Panchanathan & Robert Boyd

Center for Behavior, Evolution, and Culture and Department of Anthropology, University of California, Los Angeles, California 90095, USA

Models of large-scale human cooperation take two forms. ‘Indirect reciprocity’<sup>1</sup> occurs when individuals help others in order to uphold a reputation and so be included in future cooperation. In ‘collective action’<sup>2</sup>, individuals engage in costly behaviour that benefits the group as a whole. Although the evolution of indirect reciprocity is theoretically plausible<sup>3–6</sup>, there is no consensus about how collective action evolves. Evidence suggests that punishing free riders can maintain cooperation<sup>7–9</sup>, but why individuals should engage in costly punishment is unclear. Solutions to this ‘second-order free rider problem’ include meta-punishment<sup>10</sup>, mutation<sup>11</sup>, conformism<sup>12</sup>, signalling<sup>13–15</sup> and group-selection<sup>16–18</sup>. The threat of exclusion from indirect reciprocity can sustain collective action in the laboratory<sup>19</sup>. Here, we show that such exclusion is evolutionarily stable, providing an incentive to engage in costly cooperation, while avoiding the second-order free rider problem because punishers can withhold help from free riders without damaging their reputations. However, we also show that such a strategy cannot invade a population in which indirect reciprocity is not linked to collective action, thus leaving unexplained how collective action arises.

To show that indirect reciprocity can stabilize collective action without the second-order free rider problem, we consider a large population subdivided into randomly formed social groups of size  $n$ . Social life consists of two stages. First, individuals decide whether or not to contribute to a one-shot collective action game at a net personal cost  $C$  in order to create a benefit  $B$  shared equally amongst the  $n - 1$  other group members, where  $B > C$ . Second, individuals

engage in a multi-period ‘mutual aid game’<sup>4</sup>, a form of indirect reciprocity that is well suited to a population structured into groups. The dynamics of the mutual aid game are very similar to other models of indirect reciprocity<sup>3,6</sup> so our results should generalize to other social exchange systems. In each period of the mutual aid game, one randomly selected individual from each group is ‘needy’. Each of his  $n - 1$  neighbours can help him an amount  $b$  at a personal cost  $c$ , where  $b > c > 0$ . Each individual’s behavioural history is known to all group members. This assumption is essential because it is known that indirect reciprocity cannot evolve when information quality is poor<sup>6</sup>. The mutual aid game repeats with probability  $w$  and terminates with probability  $1 - w$ , thus lasting for  $1/(1 - w)$  periods on average. Afterwards, individuals reproduce on the basis of payoffs accumulated over both stages, relative to the whole population, and then die.

Individuals are characterized by one of three heritable strategies: ‘Defector’, ‘Cooperator’, and ‘Shunner’. Defectors do not contribute to the collective action, nor do they help during the mutual aid game. Cooperators contribute to the collective action and try to help all needy recipients during mutual aid. With probability  $e$ , however, Cooperators mistakenly fail to help recipients of good reputation in the mutual aid game owing to an implementation error<sup>6</sup> (See Box 1 for details). Shunners contribute to the collective action and then try to help those needy individuals who have good reputations during the mutual aid game, but mistakenly fail owing to errors with probability  $e$  just like Cooperators. Shunners never help needy recipients who are in bad standing.

All individuals begin their lives in good standing. Failure to contribute to the collective action results in a lifetime of bad standing. If an individual has contributed during the collective action stage, she temporarily loses her good standing if she fails to help a recipient of good reputation during the mutual aid game, either through intention or error. She can, however, restore her good standing by helping a needy recipient in some future period. Our results do not depend on the assumption that the reputations

### Box 1

#### Errors in models of reciprocity and punishment

As in previous models of indirect reciprocity<sup>3,5,6</sup>, errors play a crucial role in our analysis. These errors should not be thought of as part of an inherited strategy. Instead, they represent exogenous factors like sickness or accidents that prevent actors from helping despite an intention to do so. In our model, all group members, including the actor, know when an error has occurred. These ‘implementation’ errors are contrasted with ‘perception’ errors, in which individuals differ in their beliefs about who cooperated and who defected<sup>6</sup>. We have not analysed the effect of perception errors because these errors add sufficient mathematical complexity that analysis becomes intractable. As a result, it is unclear how perception errors affect the evolution of indirect reciprocity<sup>3,6</sup>. In addition, we do not consider errors in which individuals mistakenly help a recipient of bad reputation during the mutual aid game, nor errors during the collective action game, because both such errors complicate the model without qualitatively altering the results.

Previous models of collective action and costly punishment<sup>10–12,18</sup> have shown that implementation errors of the type we consider here undermine the evolution of collective action. To see why, suppose that there are no defecting strategies and that behaviour is error-free. In this case, selection cannot distinguish between strategies that cooperate and punish defectors and strategies that cooperate but do not punish. There is never a need to punish, so there is no second-order free rider problem. If actors occasionally defect by mistake, however, strategies that punish must do so at a personal cost. Selection will now favour strategies that cooperate but do not punish (second-order free riders). As a result, strategies that punish free riders decline and eventually defectors can invade and take over.



of collective action free riders are permanently damaged. Such an assumption does, however, simplify the analysis. The key assumption is that free-riding during the collective action stage has a more severe effect on reputation than failing to donate during a bout of reciprocity—if this were not the case, there would be little motivation to make costly contributions to the collective action. In modelling the Shunner strategy, we assume that reputation is linked across the collective action and mutual aid domains. In the next section, we examine under just what conditions selection will favour this linkage.

The model has two stable evolutionary equilibria<sup>20</sup> (Fig. 1; see Supplementary Information). Defector is always an evolutionarily stable strategy. When Shunners are common, the population resists invasion by rare Defectors if:

$$\left(\frac{n-1}{n}\right)\left(\frac{1-e}{1-w}\right)(b-c)(1-we) > C \quad (1)$$

When condition (1) is satisfied, Defectors cannot invade because the long-term benefit from mutual aid is larger than the net private cost of contributing to the collective action. Shunners can resist invasion by rare Cooperators if  $wec > 0$ , a condition that is satisfied as long as mutual aid persists, is costly, and is prone to occasional error. Selection favours Shunners over Cooperators because the two strategies behave differently towards needy individuals who failed to help a recipient of good reputation owing to error in the previous period. Cooperators always try to help these individuals at a cost  $c$ , whereas Shunners withhold aid, which is socially sanctioned because the recipient was in bad standing. In previous models of collective action with costly punishment<sup>10–12,18</sup>, errors undermine the evolution of punishment (see Box 1 for details).

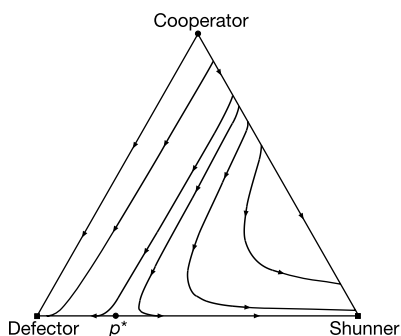
In our model, errors actually stabilize punishment because Shunners punish defection by withholding costly aid, thereby increasing their own fitness. Strategies that do not punish, such as Cooperator, are at a selective disadvantage because they dole out help indiscriminately. Note that condition (1) is not affected by the magnitude of the group benefit created by the collective action. This means that any behaviour can be stabilized by linking it to a system of indirect reciprocity, as long as the cost of that behaviour is less than the benefits derived from indirect reciprocity. Thus this process can stabilize even maladaptive norms in which  $B < C$ .

Although punishment by exclusion is evolutionarily stable in this model, so is defection (as well as another strategy, which we will consider shortly, that supports mutual aid but not collective action). Why should we believe that Shunner is the likely evolutionary

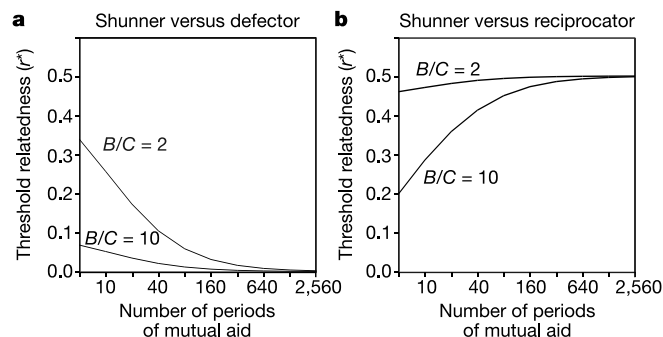
outcome? In their classic work on pair-wise reciprocity, Axelrod and Hamilton<sup>21</sup> suggest that a small amount of non-random assortment, such as interaction between relatives, destabilizes uncooperative, but not cooperative, equilibria. When pairs are formed non-randomly, a reciprocating strategy like ‘tit-for-tat’ more often channels cooperation to like types and this permits rare reciprocators to invade populations in which defecting strategies are common. To underscore the synergy between inclusive fitness and reciprocity, Axelrod and Hamilton<sup>21</sup> state that “the gear wheels of social evolution have a ratchet”. For the same reason, as we demonstrate in the Supplementary Information, low levels of assortment allow reciprocating strategies, like Shunner, to invade less cooperative strategies, like Defector, but not the reverse. As long as the information available about the behaviour of others is accurate and widespread<sup>6</sup>, there is a powerful synergy between small amounts of assortment and long periods of interaction, making increasing amounts of reciprocity the probable evolutionary outcome (Fig. 2a).

The same synergy does not, however, exist between reciprocity and collective action norms. Strategies that link the two behaviours through reputation cannot invade a population practising only indirect reciprocity. While we have demonstrated that indirect reciprocity can stabilize collective action, it is unlikely that the two behaviours evolve simultaneously as a complex. More probably, a population engaging only in indirect reciprocity (that is, mutual aid, no collective action) evolves first.

To demonstrate that mutant Shunners, who link collective action and indirect reciprocity, cannot invade a population engaged only in indirect reciprocity, despite assortment, we consider a fourth strategy, Reciprocator, which does not contribute to the collective action and also does not attend to the contributions of fellow community members. Thus, in the eyes of Reciprocators, all group members enter the stage two mutual aid game in good standing, regardless of their behaviour in the previous collective action game. Reciprocators help only good-standing recipients, assuming no error, during the mutual aid game. In their eyes, fellow group members remain in good standing as long as they help others in good standing (by their definition) during the mutual aid game. To allow for non-random interaction, we assume that the conditional probability that another individual in a group has the same strategy



**Figure 1** Evolutionary dynamics of the Shunner, Defector and Cooperator strategies, plotted in trilinear coordinates. The arrowheads depict evolutionary trajectories. Populations composed of all Shunners and all Defectors are both evolutionarily stable. When the initial frequency of Shunners is too low, the population evolves towards the asocial Defector equilibrium. When it is higher, the population reaches a stable equilibrium in which collective action is enforced by exclusion from the benefits of indirect reciprocity. The model parameters for this figure are:  $B = 10$ ,  $C = 5$ ,  $b = 2$ ,  $c = 1$ ,  $n = 100$ ,  $w = 0.95$ ,  $W_0 = 100$  and  $e = 0.05$ .



**Figure 2** The threshold degree of assortment ( $r^*$ , derived in the Supplementary Information) necessary for rare Shunners to invade as a function of the number of mutual aid periods,  $1/(1-w)$ , for two different collective action benefit cost ratios,  $B/C$ . **a**, When Shunners compete against Defectors, the assortment threshold decreases as the number of mutual aid periods increases. For long-lasting groups even small amounts of assortment allow the cooperative strategy to invade. **b**, When Shunners compete against Reciprocators the assortment threshold increases as the number of mutual aid periods increases. Rare Shunners do best in one-shot interactions. However, as shown in the text (condition (1)), Shunners are unlikely to be evolutionarily stable against Defectors in one-shot interactions. For both **a** and **b**, the parameter values are  $b = 2$ ,  $c = 1$ ,  $B = 10$ ,  $n = 30$  and  $e = 0.01$ . The graph plotted in **b** is based on a compete condition derived in the Supplementary Information that does not depend on the assumption that  $r$  is small.

as a focal individual is given by  $r + (1 - r)p$  where  $p$  is the frequency of that strategy in the population and  $0 \leq r \leq 1$ . Thus the parameter  $r$  is closely analogous to the familiar coefficient of relatedness<sup>22</sup>. When  $r = 0$ , groups are formed at random. Larger values of  $r$  mean that individuals tend to interact with others of the same strategy. Setting the frequency of Cooperator and Defector to zero, we can now ask whether rare Shunners can invade a population in which Reciprocators are common.

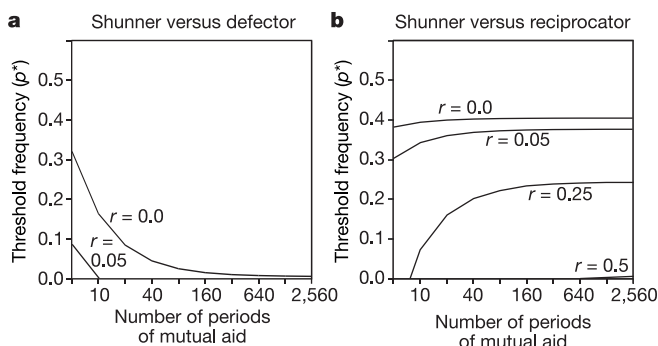
Small amounts of non-random assortment do not allow Shunner to invade a population in which Reciprocators are common, even if interactions go on indefinitely (Fig. 2b). When groups are formed at random ( $r = 0$ ), both Shunner and Reciprocator are evolutionarily stable strategies. The basin of attraction for Shunner becomes smaller as the number of periods of mutual aid increases (Fig. 3b). As  $r$  is increased, a threshold value,  $r^*$ , is eventually reached; above this value Shunners always eliminate Reciprocators. However, increasing the number of mutual aid periods increases the value of  $r^*$  (Fig. 2b). This means that Shunners increase when initially rare only if collective action increases inclusive fitness in one-shot interactions<sup>22</sup> (that is,  $w \approx 0$  and  $rB > C$ )—assortment and reciprocity now operate antagonistically. Axelrod and Hamilton's ratchet is nowhere in evidence.

To see why, consider condition (2) which must be met for Shunners to increase when rare:

$$r \left[ B + \left( \frac{1}{1-w} \right) (2bw - c) \right] > C + \left( \frac{1}{1-w} \right) (wb - c) \quad (2)$$

Mutual aid from Shunners                      Mutual aid from Reciprocators

(In deriving this expression, we assume that  $e = 0$ ,  $n$  is large, and  $r$  is sufficiently small that terms of order  $r^2$  can be ignored. These



**Figure 3** The threshold frequency at which Shunners increase ( $p^*$ ) as a function of the number of mutual aid periods,  $1/(1 - w)$ , for different levels of assortment,  $r$ . If the initial frequency of Shunner is greater than  $p^*$ , Shunner increases in frequency, eventually eliminating the competing strategy. If the initial frequency of Shunner is less than  $p^*$ , Shunner is eliminated. Shunner cannot coexist with either Defector or Reciprocator in a stable polymorphic equilibrium state. **a**, When Shunner competes against Defector, the basin of attraction for Shunner becomes larger (that is,  $p^*$  decreases) as the number of periods of mutual aid increases. Increasing assortment amplifies this effect such that Shunner may become the only evolutionarily stable strategy (that is,  $p^* \leq 0$ ). **b**, In contrast, when Shunner competes against Reciprocator, the basin of attraction for Shunner decreases as the number of mutual aid periods increases, for all levels of assortment. Increasing assortment increases the basin of attraction for Shunner. The Reciprocator equilibrium can be destabilized (that is,  $p^* \leq 0$ ) when assortment is high and there are few bouts of mutual aid (for example,  $r = 0.25$  and  $w < 0.9$ ). Under such conditions, however, Shunner is not evolutionarily stable against Defector (see condition (1)). When mutual aid is long-lasting, the Reciprocator equilibrium can still be destabilized only if assortment is very high (for example, if  $r = 0.5$ ). Further analysis indicates that when condition (1) is satisfied and the collective action is beneficial for the group ( $B > C$ ), the Shunner strategy has a larger basin of attraction than Reciprocator ( $p^* < 0.5$ ) which means that within-group equilibrium selection processes<sup>23–27</sup> will favour the Shunner strategy. The parameter values for both **a** and **b** are  $b = 2$ ,  $c = 1$ ,  $C = 1$ ,  $B = 10$ ,  $n = 30$  and  $e = 0.01$ .

simplifications have no qualitative effect.) Notice that there are terms that represent the benefits of mutual aid on both sides of the inequality. The term on the benefit (left) side gives the long-term benefits of mutual aid from other Shunners, whereas the term on the cost (right) side gives the forgone mutual aid benefits from Reciprocators. In the first stage of social interaction, Shunners contribute to the collective action whereas Reciprocators do not. On entering stage two, Reciprocators consider all others to be in good standing, whereas Shunners consider only other Shunners to be worthy. Shunners permanently impugn Reciprocators for their failure to contribute during stage one. Shunners therefore do not aid Reciprocators in the mutual aid game and in turn fall into bad standing in the eyes of Reciprocators. Thus, when a Shunner needs help in stage two, few Reciprocators oblige. As a result, Shunners lose out on most of the benefits of mutual aid. Differing opinions of good citizenship and impropriety have driven a moral wedge into the community. In the Supplementary Information, we also consider a situation, different from the one presented, in which a few individuals attempt to change an existing social norm. That is, we assume that the population is playing the Shunner strategy and a few mutants attempt to change from the current collective action to another one (for example, forest clearing to forest preservation). The results are similar to those just presented; assortment and reciprocity operate antagonistically.

That indirect reciprocity can stabilize collective action is nonetheless significant. In other models<sup>10,12,16,18</sup>, collective action is stabilized by direct punishment in which individuals pay a private cost in order to inflict a still greater cost on free riders. Such punishment is undermined by the presence of second-order free riders which ultimately destabilizes cooperation. When collective action free riders are indirectly punished through exclusion from ongoing social exchange (for example, indirect reciprocity) there is no second-order free rider problem. Instead of bearing a private cost, individuals benefit by withholding aid. The group pays the cost when individuals engage in such indirect punishment because withholding aid from one community member lowers mean fitness.

Where then does this leave us with respect to the evolution of collective action through indirect reciprocity? There are a number of equilibrium selection processes that could explain the transition from a population engaging in pure indirect reciprocity (such as Reciprocator) to one linking indirect reciprocity to a collective action (such as Shunner). Within-group processes based on random fluctuations<sup>23–26</sup> or individual calculation<sup>27</sup> typically pick out the equilibrium with the largest basin of attraction. Between-group processes that result from inter-group competition<sup>12,28,29</sup> or the diffusion of ideas from more successful groups to less successful ones<sup>30</sup> promote the systematic spread of strategies that lead to higher average group payoff. All of these processes favour the Shunner strategy over Reciprocator. □

Received 12 May; accepted 28 August 2004; doi:10.1038/nature02978.

- Alexander, R. D. *The Biology of Moral Systems* (de Gruyter, New York, 1987).
- Olson, M. *The Logic of Collective Action: Goods and the Theory of Groups* (Harvard Univ. Press, Cambridge, 1971).
- Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574 (1998).
- Sugden, R. *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford, 1986).
- Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268**, 2495–2501 (2001).
- Panchanathan, K. & Boyd, R. A tale of two defectors: the importance of standing for the evolution of reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).
- Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116 (1986).
- Ostrom, E. J., Walker, J. & Gardner, R. Covenants with and without a sword: self-governance is possible. *Am. Polit. Sci. Rev.* **86**, 404–417 (1992).
- Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
- Boyd, R. & Richerson, P. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
- Sethi, R. & Somanathan, E. The evolution of social norms in common property resource use. *Am. Econ. Rev.* **86**, 766–788 (1996).
- Henrich, J. & Boyd, R. Why people punish defectors—Weak conformist transmission can stabilize

- costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* **208**, 79–89 (2001).
13. Smith, E. A. & Bliege Bird, R. L. Turtle hunting and tombstone opening: public generosity as costly signaling. *Evol. Hum. Behav.* **21**, 245–261 (2000).
  14. Gintis, H., Smith, E. A. & Bowles, S. Costly signaling and cooperation. *J. Theor. Biol.* **213**, 103–119 (2001).
  15. Roberts, G. Competitive altruism: from reciprocity to the handicap principle. *Proc. R. Soc. Lond. B* **265**, 427–431 (1998).
  16. Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179 (2000).
  17. Sober, E. & Wilson, D. S. *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Harvard Univ. Press, Cambridge, 1999).
  18. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535 (2003).
  19. Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).
  20. Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge, 1982).
  21. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
  22. Hamilton, W. D. The genetical evolution of social behavior. Parts I, II. *J. Theor. Biol.* **7**, 1–52 (1964).
  23. Kandori, M., Mailath, G. & Rob, R. Learning, mutation, and long-run equilibria in games. *Econometrica* **61**, 29–56 (1993).
  24. Young, P. H. The evolution of conventions. *Econometrica* **61**, 57–84 (1993).
  25. Ellison, G. Learning, local interaction, and coordination. *Econometrica* **61**, 1047–1071 (1993).
  26. Samuelson, L. *Evolutionary Games and Equilibrium Selection (Economic Learning and Social Evolution)* (MIT Press, Cambridge, 1997).
  27. Bowles, S. *Microeconomics: Behavior, Institutions, and Evolution* (Princeton Univ. Press, Princeton, 2004).
  28. Boyd, R. & Richerson, P. Group selection among alternative evolutionary stable strategies. *J. Theor. Biol.* **145**, 331–342 (1990).
  29. Soltis, J., Boyd, R. & Richerson, P. Can group-functional behaviors evolve by cultural group selection—An empirical test. *Curr. Anthropol.* **36**, 473–494 (1995).
  30. Boyd, R. & Richerson, P. Group beneficial norms spread rapidly in a structured population. *J. Theor. Biol.* **215**, 287–296 (2002).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank C. Barrett, S. Bowles, E. Fehr, H. Gintis, J. Henrich, R. Kurzban, S. Naficy, C. Navarrete, D. Penn, J. Silk and the Experimental Biological Anthropology Group for discussions. We especially thank K. Haley for developing the ideas presented in this paper. K.P. was funded by an NSF Graduate Research Fellowship.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to K.P. ([buddha@ucla.edu](mailto:buddha@ucla.edu)).

## Physical performance and darwinian fitness in lizards

Jean-François Le Galliard<sup>1\*</sup>, Jean Clobert<sup>1</sup> & Régis Ferrière<sup>1,2</sup>

<sup>1</sup>Laboratoire Fonctionnement et Evolution des Systèmes Ecologiques, CNRS UMR 7625, Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris cedex 05, France

<sup>2</sup>Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

\* Present address: Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, P.O. Box 1050, Blindern, 0316 Oslo, Norway

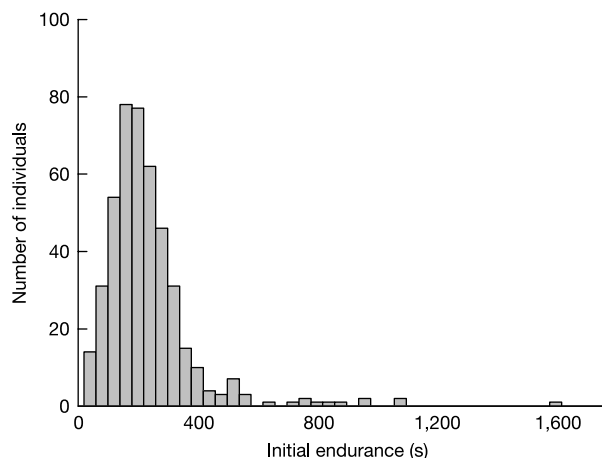
Strong evidence for a genetic basis of variation in physical performance has accumulated<sup>1,2</sup>. Considering one of the basic tenets of evolutionary physiology—that physical performance and darwinian fitness are tightly linked<sup>3</sup>—one may expect phenotypes with exceptional physiological capacities to be promoted by natural selection. Why then does physical performance remain considerably variable in human and other animal populations<sup>1,2,4</sup>? Our analysis of locomotor performance in the common lizard (*Lacerta vivipara*) demonstrates that initial endurance (running time to exhaustion measured at birth) is indeed highly heritable, but natural selection in favour of this trait can be unexpectedly weak. A manipulation of dietary conditions unravels a proximate mechanism explaining this pattern. Fully fed individuals experience a marked reversal of

performance within only one month after birth: juveniles with low endurance catch up, whereas individuals with high endurance lose their advantage. In contrast, dietary restriction allows highly enduring neonates to retain their locomotor superiority as they age. Thus, the expression of a genetic predisposition to high physical performance strongly depends on the environment experienced early in life.

Sporting events would be exceedingly boring were there no variation in human performance; fortunately, this is not the case. For example, the distribution of finish times at international marathons has a large variance and a long tail<sup>1</sup>, due to a variety of factors affecting the performance of individual runners<sup>5</sup>. Although genetic variation in locomotor performance has been documented in human and other animal populations<sup>1,2</sup>, questions remain as to how genetic and non-genetic factors would interact with each other and what effect selection has on the resulting individual variation<sup>1,6</sup>. We addressed these two questions using ground-dwelling lizards, a popular model system for studies of locomotor performance<sup>2,4,7</sup>. Our focus here is on the endurance capacity as assayed in the laboratory (see Methods). In lizards, endurance shows considerable interindividual variation that reflects differences in tight muscle mass, heart mass and aerobic metabolism<sup>8</sup>.

Our study species is the common lizard (*Lacerta vivipara* Jacquin 1787) for which locomotor performance and life-history traits have been routinely studied<sup>9</sup>. We took advantage of the populations established at the Ecological Research Station of Foljuif (Nemours, France) in the semi-natural conditions of outdoor enclosures<sup>10</sup> to measure the heritability of initial endurance and the age-specific strength of natural selection on this trait. In these enclosures, endurance capacity could reflect social rank<sup>2</sup> and abilities to compete for and exploit basking sites and prey<sup>4</sup>, and thus influence darwinian fitness<sup>9</sup>. Insights into proximate mechanisms underlying the observed pattern of selection have been gained experimentally by investigating how dietary conditions early in life influence the ontogeny of endurance and the relationship between survivorship and endurance.

In 2001, initial endurance was recorded in a cohort of 447 offspring (Fig. 1). Measurements spanned a 45-fold range, from 36 s to 1,677 s (mean = 222 s ± 153.7 s.d.). The distribution is typically skewed, with a few 'champions' displaying exceptional endurance. Initial endurance increased with offspring body size and body condition, decreased with maternal body size, and increased with behavioural motivation (Table 1). Accounting for all these factors, initial endurance was highly heritable ( $h^2 = 0.40$ ), concurring with previous studies in this species and many other reptiles<sup>2,11</sup>. Even in the controlled conditions of our outdoor enclosures, high heritability might have been caused by maternal effects, but no such



**Figure 1** Individual variation in endurance capacity among 447 common lizard offspring.



Table 1 Proximate factors of initial endurance

|                          | Test statistics            | Parameter estimates ( $\pm$ s.e.) | Partial regression $r^2$ |
|--------------------------|----------------------------|-----------------------------------|--------------------------|
| Fixed effects            |                            |                                   |                          |
| Offspring body size      | $F_{1,355} = 14.6^*$       | $0.098 \pm 0.026$                 | 0.041                    |
| Offspring body condition | $F_{1,355} = 40.4^\dagger$ | $8.725 \pm 1.373$                 | 0.070                    |
| Behavioural motivation   | $F_{1,355} = 86.1^\dagger$ | $-0.031 \pm 0.012$                | 0.233                    |
| Maternal body size       | $F_{1,355} = 7.1^*$        | $-0.953 \pm 0.103$                | 0.046                    |
| Random effects           |                            |                                   |                          |
| Family                   | $\chi^2 = 105.1^\dagger$   | $0.125 \pm 0.026$                 | —                        |

Behavioural motivation was measured by the number of stimulations per unit distance. Endurance was log-transformed to ensure normality and homoscedasticity. The final model was obtained after a stepwise multivariate linear regression. Fixed factors included offspring characteristics, that is body size (snout-vent length), residual tail size (relative to body size), condition (residual body mass) and sex; maternal characteristics, that is body size, residual fecundity (relative to maternal body size) and post-partum body condition. Family was included as a random effect, and the broad-sense heritability was calculated from variance components<sup>27</sup>. Statistical tests are type III  $F$  tests for fixed effects and likelihood ratio tests for random effects.

\* $P < 0.01$ ,  $^\dagger P < 0.001$ .

effect (maternal age or population density, see Methods) was detected (analysis of variance (ANOVA) with family included as a random factor nested within age and density effects; age effect:  $F_{2,45} = 0.88$ , NS; density effect:  $F_{1,45} = 0.10$ , NS; age  $\times$  density effect:  $F_{2,45} = 0.17$ , NS).

Juvenile survival is an important component of darwinian fitness in the common lizard<sup>12</sup>. Among the 447 offspring scored at birth, 16 individuals died before release, 316 survived during the summer and 192 survived after one year. Natural selection acted independently on morphology and endurance (non-significant patterns of correlational selection among body size, body condition and endurance;  $P > 0.13$ ). Directional selection for greater endurance and larger body size at birth was detected over the first summer and over the whole year following birth, with most selection taking place shortly after birth (Table 2; end of summer to next year selection on initial endurance:  $F_{1,234} = 0.01$ , NS; on initial body size:  $F_{1,234} = 2.86$ , NS). However, juvenile survival selection on endurance was highly sensitive to the few lizards with lowest initial endurance (Table S1 in Supplementary Information), which agrees with the shape of the fitness function (Fig. 2). Thus, natural selection acted predominantly against very low initial endurance and was nearly neutral at intermediate and high levels of endurance.

Weak selection for elite endurance is at odds with the common assumption that performance and darwinian fitness are tightly correlated. One explanation for our findings might have been that endurance reflected motivational factors in the laboratory more than physiological capacities (Table 1), but behavioural motivation did not correlate with survival probabilities, and factoring motivation out of the selection analyses had no effect on our main results (Table S2 in Supplementary Information). Moreover, the pattern of selection demonstrated by this experiment is concordant with two previous correlative studies of initial endurance and survival in

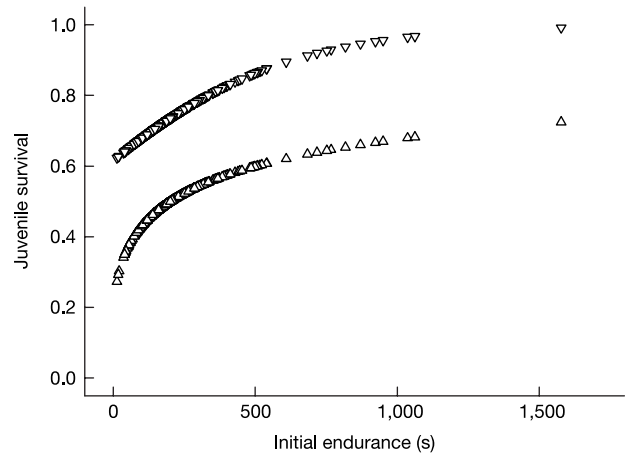


Figure 2 Fitness functions of initial endurance during the first summer (triangles pointing down) and the whole year (triangles pointing up) following birth. Data have been back-transformed from values predicted by the models (see Table 2). Each triangle corresponds to a single individual.

natural populations<sup>9,13</sup>. Our analysis also found a strong positive effect of selection on body size at birth, which is similar to the effect detected in natural populations of the same species<sup>14</sup>, and in other vertebrates<sup>15</sup>. The overall coherence of these observations suggests that our results are not an artefact of measuring selection in the semi-natural conditions of our enclosures (see Discussion in Supplementary Information).

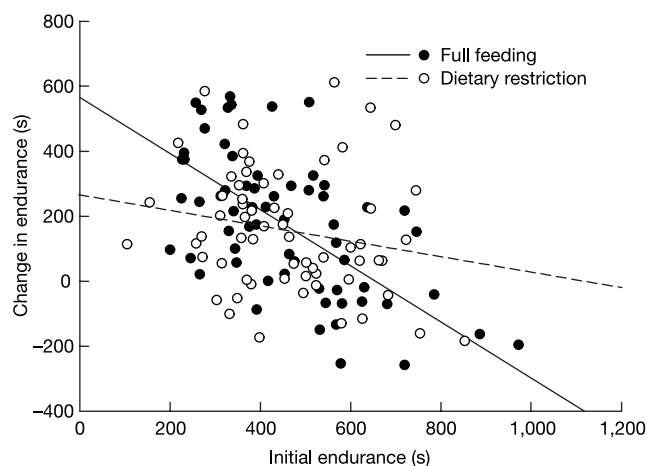
In 2002, we tested the hypothesis that rapid ontogenic shifts in endurance under favourable conditions could explain this unexpected pattern of selection. We measured the change in endurance over the month following birth in two samples subjected to different dietary conditions, and then assessed the relationship between initial endurance and annual survival by releasing these individuals in two outdoor enclosures (see Methods). The treatments were carefully calibrated to mimic full feeding and dietary restriction compared with the favourable conditions of our enclosures (see Table S3 in Supplementary Information). Overall, individuals gained endurance as they grew up (paired Student's  $t$ -test,  $P < 0.001$ ; average individual change:  $158 \text{ s} \pm 19.4 \text{ s.e.}$ ). The ontogenic change in endurance was affected by an interaction between diet and initial endurance ( $F_{1,90} = 10.9$ ,  $P < 0.01$ ): under dietary restriction, initial endurance had no effect on the change in endurance ( $F_{1,28} = 1.34$ , NS); in contrast, the performance of fully fed juveniles with low initial endurance rose markedly, whereas the performance of individuals with high initial performance dropped ( $F_{1,29} = 58.9$ ,  $P < 0.001$ ; Fig. 3). Thus, individual differences in endurance were

Table 2 Natural selection on initial endurance, body size and body condition

|                       | Summer survival            |                         |          | Annual survival            |                         |          |
|-----------------------|----------------------------|-------------------------|----------|----------------------------|-------------------------|----------|
|                       | Test statistics            | Estimates ( $\pm$ s.e.) | Gradient | Test statistics            | Estimates ( $\pm$ s.e.) | Gradient |
| Fixed effects         |                            |                         |          |                            |                         |          |
| Endurance (linear)    | $F_{1,344} = 6.42^\dagger$ | $0.465 \pm 0.183$       | 0.104    | $F_{1,343} = 4.43^\dagger$ | $0.268 \pm 0.127$       | 0.120    |
| Body size (linear)    | $F_{1,344} = 7.17^\dagger$ | $0.341 \pm 0.127$       | 0.077    | $F_{1,343} = 7.45^\dagger$ | $0.335 \pm 0.123$       | 0.150    |
| Condition (linear)    | $F_{1,344} = 3.3^*$        | $0.249 \pm 0.137$       | 0.056    | $F_{1,343} = 1.27$         | $0.146 \pm 0.130$       | 0.065    |
| Condition (quadratic) | NS                         | NS                      | NS       | $F_{1,343} = 5.92^\dagger$ | $0.228 \pm 0.094$       | 0.102    |
| Random effects        |                            |                         |          |                            |                         |          |
| Enclosure             | $\chi^2 = 39.39^\S$        | 0.491                   | —        | $\chi^2 = 60.83^\S$        | 0.743                   | —        |
| Family                | $\chi^2 = 52.56^\S$        | 0.531                   | —        | $\chi^2 = 52.63^\S$        | 0.441                   | —        |

Natural selection was studied over the summer following birth (summer survival), and over the first year of life (annual survival). Endurance, body size and body condition were standardized (zero mean, unit variance) for 431 lizards (9 enclosures, 84 families). Survival probability was modelled with mixed-effects logistic regressions using the GLIMMIX macro in SAS<sup>28</sup>. The final models were obtained after stepwise multivariate analyses. We used linear terms to test for patterns of directional selection on single traits, quadratic terms to test for patterns of stabilizing or disruptive selection on single traits, and mixed polynomial terms to test for correlational selection on pairs of traits. Enclosure effects and family effects nested within enclosures were modelled as random effects. Models adequately fitted the data<sup>29</sup>, and qualitatively matched non-parametric, cubic spline regressions of fitness functions<sup>29</sup>. Parameter estimates are given on a logistic scale. Standardized selection gradients are obtained from the slope terms of the logistic regression<sup>30</sup>.

\* $0.05 < P < 0.10$ ,  $^\dagger P < 0.05$ ,  $^\S P < 0.01$ ,  $^\P P < 0.001$ .

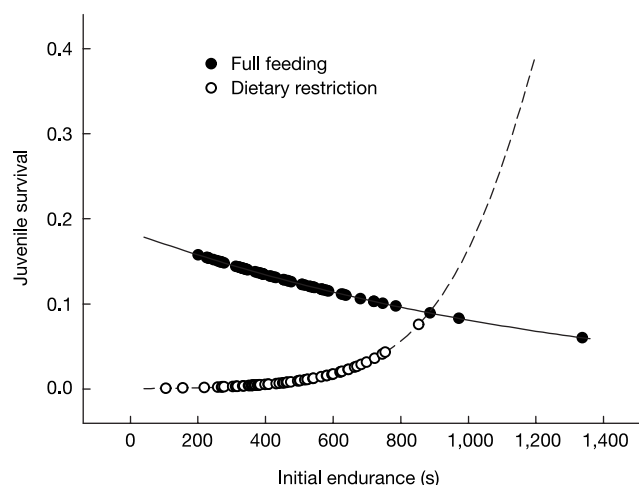


**Figure 3** Ontogenic change in endurance depending on diet and initial endurance. Straight lines are regression lines in full feeding conditions (solid line) and under dietary restriction (dashed line). The one-month change in endurance was modelled as a function of diet, sex, initial endurance and interactions between these factors. Blocks (trays, see Methods) and family within blocks were treated as random effects. The Pearson product-moment correlation coefficient between initial endurance and endurance at the age of one month was higher under dietary restriction ( $r = 0.57$ ) than in the full feeding treatment ( $r = 0.21$ , Student's  $t$ -test on  $z$ -transformation,  $P = 0.018$ ).

consistent across ontogeny only under dietary restriction (ANOVA on subject effect,  $F_{63,64} = 2.17$ ,  $P < 0.01$ , intra-class correlation coefficient  $r = 0.37$ ; with rich diet,  $F_{62,63} = 0.99$ , NS). We therefore expected that after releasing both groups of lizards in field enclosures, the effect of natural selection on initial endurance would be stronger in the dietary restricted group than in the fully fed group, and found support for this prediction (Fig. 4). Thus, dietary conditions experienced early in life influence both the ontogenic consistency of endurance and the predictability of natural selection on variation in initial endurance.

Previous evidence of ontogenic consistency of physical performance was mainly restricted to post-growth life-history stages<sup>4</sup>, and as far as we know, no study so far has tested the hypothesis that ontogenic consistency could be sensitive to environmental conditions. The pattern of ontogenic change reported here suggests that low- and high-performance juveniles utilize different resource allocation strategies when resources are abundant. Low-performance individuals might allocate resources to muscles and aerobic metabolism resulting in enhanced locomotor performance, whereas high-performance individuals would either reallocate resources towards growth and maturation, or direct more energy to fat reserves with impaired locomotion as a side effect<sup>16,17</sup>. Our finding of a positive correlation between change in endurance and growth in body size under full feeding conditions ( $F_{1,29} = 7.37$ ,  $P = 0.01$ ), whereas no correlation arose under dietary restriction ( $F_{1,28} = 1.71$ , NS), lends weight to this allocation trade-off hypothesis.

Evolutionary physiologists have assumed ontogenic consistency of locomotor traits and a strong, positive relationship between locomotor performance and darwinian fitness<sup>3</sup>. Our results challenge these basic tenets of evolutionary physiology: ontogenic consistency depends upon environmental conditions, thus limiting the predictability of natural selection on performance at birth. Variation in food availability occurs in wild populations of the common lizard<sup>18</sup>; our experimental results predict that natural selection on initial endurance is ineffective in high-food years or locations. Under unfavourable conditions, other mechanisms such as behavioural compensations (for example, active versus sit-and-wait foraging strategies<sup>7</sup>) or trade-offs (for example, differential exposure to predators or parasites<sup>9</sup>) might also weaken selection on initial endurance. Lack of ontogenic consistency should further



**Figure 4** Fitness functions of initial endurance depending on dietary conditions experienced during the first month following birth (filled circles and solid line, full feeding; open circles and dashed line, dietary restriction). Data have been back-transformed from values predicted by the logistic regression; each circle corresponds to a single individual. The annual survival probability was significantly affected by dietary conditions ( $F_{1,92} = 7.29$ ,  $P = 0.0083$ ) and by an interaction between dietary conditions and initial endurance ( $F_{1,92} = 2.98$ , one-tailed test of the directional hypothesis,  $P = 0.04$ ), while controlling for differences between enclosures and amongst families.

cause low heritability of endurance at the yearling and adult stages, and therefore contribute to the maintenance of individual variation for physical performance at all life-history stages.

There are other contexts in which the paradox arises of a trait measured at birth being a positive influence of darwinian fitness early in life, but having little effect on fitness at later stages or over the whole life<sup>15,19</sup>. Persistently high heritability for such a trait was previously explained by changes during the life cycle in how selection operates<sup>19</sup>. Our study emphasizes that locomotor traits are embedded in a more complex, dynamic phenotype, and provides evidence for the role of ontogeny to loosen the link between the initial value of the trait and darwinian fitness<sup>20</sup>. As a consequence, specific conditions applied early in life would seem necessary to counter developmental effects on endurance and to ensure the expression of a genetic predisposition to high physical performance. □

## Methods

### Species

The common lizard is viviparous (modal clutch size: 5–6 eggs) and offspring are autonomous at birth. Lizards used in this study were monitored in enclosed populations located at the Ecological Research Station of Foljuif (60 m above sea level, 48°17' N, 2°41' E). Our selection study avoids the limitations of previous works<sup>2</sup> by using multivariate selection analysis on both performance and morphological traits<sup>3</sup>; by controlling survival estimates for capture and movement heterogeneity; and by using outdoor enclosures to limit environmental heterogeneity.

### Maternal effects

To evaluate the effects of maternally experienced population density on offspring performance, a sub-sample of our populations (257 offspring from 51 families) was manipulated in 1999 so that reproductive females experienced two levels of population density<sup>10</sup>. During the gestation interval for the young born in 2001, the population size was 9.6 individuals ( $\pm 3.6$  s.e.) in the low-density and 21.3 individuals ( $\pm 4.4$  s.e.) in the high-density enclosures ( $\chi^2 = 9.0$ ,  $P < 0.01$ ). We also defined three age classes of mothers: 2-year-old, 3-year-old, and older.

### Sampling protocol

In June 2001, 89 gravid females were removed from outdoor enclosures and maintained in the laboratory. Females were measured (snout–vent length) and weighed regularly during gestation. After parturition, females were weighed and offspring were sexed, sized (snout–vent length, tail length, mass), individually marked (toe-clipping) and isolated in individual terraria. Endurance was measured on the day following birth.

## Endurance trials

Endurance was measured on a circular treadmill<sup>21</sup>. Lizards (warmed up and maintained at a temperature close to their field optimum) were stimulated to run at a constant speed by gently tapping the base of their tail with a soft paintbrush. Endurance capacity was measured as the time to exhaustion (to the closest second), signalled by the lack of response after 10 consecutive taps<sup>22</sup>. There was a highly significant correlation between first and second measurements ( $r = 0.78$ ) (ANOVA on 70 offspring from 14 families measured two days apart, log-transformed endurance,  $F_{69,70} = 10.18$ ,  $P < 0.001$ ).

## First analysis of natural selection

In July 2001, all offspring were released in nine enclosures, each receiving 10 families, 21 adults (including mothers) and 16 yearlings. Two recapture sessions took place in August 2001 (average age = 33 days  $\pm$  7.7 s.d.) and June 2002 (average age = 311 days  $\pm$  9.9 s.d., all individuals were then removed to the laboratory). Capture probabilities in August 2001 were estimated by fitting probabilistic models of individual capture-recapture histories<sup>23</sup>. Estimates were very close to one, allowing us to assume that individuals not seen in August 2001 had died before that census.

## Dietary effects on ontogenic consistency and second analysis of natural selection

In 2002, we performed a laboratory manipulation of rations during the four weeks following birth. Two food treatments were designed on the basis of our unpublished growth data (used to calibrate a realistic distribution of postnatal growth rates), and physiological data<sup>24</sup> (to translate growth rates into expected food intakes in the laboratory). Dietary restriction was set to a delivery of 15 mg day<sup>-1</sup> of house cricket larvae (*Acheta domestica*, 3–5 mm size) during the first week, and raised each subsequent week to match the pattern of individual growth (see Table S4 in Supplementary Information). The full feeding treatment followed a parallel pattern in a 1:3 ratio. Over the entire manipulation, average food provision was 20.5 mg day<sup>-1</sup> in the low-food treatment and 61.5 mg day<sup>-1</sup> in the high-food treatment. To compare siblings, we selected two males and two females from 32 families and allocated one individual of each sex to each treatment. We recorded endurance at the age of one day and at the end of the manipulation (age 33 days). Siblings were then released at random in one of two outdoor enclosures where populations of 13 adults, 30 yearlings and 10 juveniles had been established two months earlier. All individuals were removed from the enclosures in late May of the following year. The difference in annual survival probabilities between this and the first analysis of natural selection is likely to reflect costs of settlement in already populated enclosures and costs of translocation from the laboratory<sup>25</sup>. The directional prediction that dietary conditions should affect the relationship between endurance and survival probability was tested with a mixed-effects logistic regression using a one-tailed test<sup>26</sup>.

Received 29 July; accepted 22 September 2004; doi:10.1038/nature03057.

- Rupert, J. L. The search for genotypes that underlie human performance phenotypes. *Comp. Biochem. Physiol. A* **136**, 191–203 (2003).
- Garland, T. J. & Losos, J. in *Ecological Morphology: Integrative Organismal Biology* (eds Wainwright, P. C. & Reilly, S. M.) 240–302 (Univ. Chicago Press, Chicago, 1994).
- Arnold, S. J. Morphology, performance and fitness. *Am. Zool.* **23**, 347–361 (1983).
- Bennett, A. F. & Huey, R. B. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. J. & Antonovics, J.) 251–284, (1990).
- Bouchard, C., Malina, R. M. & Péresse, L. *Human Kinetics* 408 (Champaign, Illinois, 1997).
- Irschick, D. J. Evolutionary approaches for studying functional morphology: examples from studies of performance capacity. *Integr. Comp. Biol.* **42**, 278–290 (2002).
- Irschick, D. J. & Garland, T. J. Integrating function and ecology in studies of adaptation: investigations of locomotor capacity as a model system. *Annu. Rev. Ecol. Syst.* **32**, 367–396 (2001).
- Garland, T. J. Physiological correlates of locomotor performances in a lizard: an allometric approach. *Am. J. Physiol.* **247**, R806–R815 (1984).
- Clobert, J. *et al.* Trade-offs in phenotypic traits: endurance at birth, growth, survival, predation and susceptibility to parasitism in a lizard, *Lacerta vivipara*. *Funct. Ecol.* **14**, 675–684 (2000).
- Le Galliard, J.-F., Ferrière, R. & Clobert, J. Mother-offspring interactions affect natal dispersal in a lizard. *Proc. R. Soc. Lond. B* **270**, 1163–1169 (2003).
- Sorci, G., Swallow, J. G., Garland, T. J. & Clobert, J. Quantitative genetics of locomotor speed and endurance in the lizard *Lacerta vivipara*. *Physiol. Zool.* **68**, 698–720 (1995).
- Lorenzon, P., Clobert, J. & Massot, M. The contribution of phenotypic plasticity to adaptation in *Lacerta vivipara*. *Evolution* **55**, 392–404 (2001).
- Jayne, B. C. & Bennett, A. F. Selection of locomotor performance capacity in a natural population of garter snakes. *Evolution* **44**, 1204–1229 (1990).
- Sorci, G. & Clobert, J. Natural selection on hatchling body size and mass in two environments in the common lizard (*Lacerta vivipara*). *Evol. Ecol. Res.* **1**, 303–316 (1999).
- Janzen, F. J., Tucker, J. K. & Paukstis, G. L. Experimental analysis of an early life-history stage: selection on size of hatchling turtles. *Ecology* **81**, 2290–2304 (2000).
- Cureton, K. J. & Sparling, P. B. Distance running performance and metabolic responses to running in men and women with excess weight experimentally equated. *Med. Sci. Sports Exerc.* **12**, 288–294 (1980).
- Swallow, J. G., Koteja, P., Carter, P. A. & Garland, T. Jr. Food consumption and body composition in mice selected for high wheel-running activity. *J. Comp. Physiol. B* **171**, 651–659 (2001).
- Khodadoost, M., Pilorge, T. & Ortega, A. Variations de la densité et de la taille corporelle en fonction de la composition du peuplement de proies de trois populations de lézards vivipares du Mont Lozère. *Revue d'Ecologie (Terre Vie)* **42**, 193–201 (1987).
- Price, T. D. & Grant, P. R. Life history traits and natural selection for small body size in a population of Darwin's finches. *Evolution* **38**, 483–494 (1984).
- Riska, B. Composite traits, selection response, and evolution. *Evolution* **43**, 1172–1191 (1986).
- Le Galliard, J.-F., Le Bris, M. & Clobert, J. Timing of locomotor impairment and shift in thermal preferences during gravidity in a viviparous lizard. *Funct. Ecol.* **17**, 877–885 (2003).

- Huey, R. B., Bennett, A. F., John-Alder, H. & Nagy, K. A. Locomotor capacity and foraging behaviour of Kalahari lacertid lizards. *Anim. Behav.* **32**, 41–50 (1984).
- White, G. C. & Burnham, K. P. Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**, 120–138 (1999).
- Avery, R. A. Estimates of food consumption by the lizard *Lacerta vivipara* Jacquin. *J. Anim. Ecol.* **40**, 351–365 (1971).
- Massot, M., Clobert, J., Lecomte, J. & Barbault, R. Incumbent advantage in common lizards and their colonizing ability. *J. Anim. Ecol.* **63**, 431–440 (1994).
- Rice, W. R. & Gaines, S. D. 'Heads I win, tails you lose': testing directional alternative hypotheses in ecological and evolutionary research. *Trends Ecol. Evol.* **9**, 235–237 (1994).
- Falconer, D. S. *An Introduction to Quantitative Genetics* (Longman, London, 1989).
- Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. *SAS Systems for Mixed Models* (SAS Institute, Cary, North Carolina, 1996).
- Schluter, D. Estimating the form of natural selection on a quantitative trait. *Evolution* **42**, 849–861 (1988).
- Janzen, F. J. & Stern, H. S. Logistic regression for empirical studies of multivariate selection. *Evolution* **52**, 1564–1571 (1998).

Supplementary Information accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We are grateful to A. Badyaev, A. Burke and T. Markow for comments, to P. Cassey, M. Massot, M. Nachman, D. Papaj, M. Saunders and T. van Dooren for discussions, and to J. Cote, B. Decencière, M.-L. Jarzat, M. Le Bris, D. Mersch and S. Testard for assistance. Financial support was received from the French Ministry of Education and Research, the French Ministry of Environment, the European Research Training Network 'ModLife', and the NSF Biomath REU programme.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to J.-F.L.G. (j.f.l.galliard@bio.uio.no).

# Predator-induced behaviour shifts and natural selection in field-experimental lizard populations

Jonathan B. Losos<sup>1</sup>, Thomas W. Schoener<sup>2</sup> & David A. Spiller<sup>2</sup>

<sup>1</sup>Department of Biology, Campus Box 1137, Washington University, St Louis, Missouri 63130, USA

<sup>2</sup>Section of Ecology and Evolution and Center for Population Biology, University of California, Davis, California 95616, USA

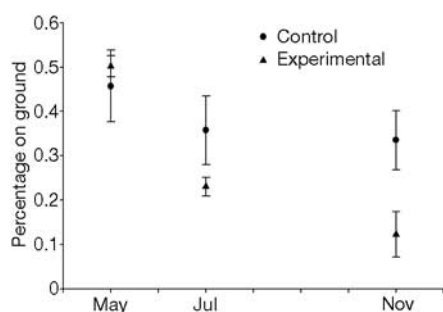
The role of behaviour in evolutionary change has long been debated. On the one hand, behavioural changes may expose individuals to new selective pressures by altering the way that organisms interact with the environment, thus driving evolutionary divergence<sup>1–3</sup>. Alternatively, behaviour can act to retard evolutionary change<sup>4–6</sup>: by altering behavioural patterns in the face of new environmental conditions, organisms can minimize exposure to new selective pressures. This constraining influence of behaviour has been put forward as an explanation for evolutionary stasis within lineages<sup>4,7–9</sup> and niche conservatism within clades<sup>10,11</sup>. Nonetheless, the hypothesis that behavioural change prevents natural selection from operating in new environments has never been experimentally tested. We conducted a controlled and replicated experimental study of selection in entirely natural populations; we demonstrate that lizards alter their habitat use in the presence of an introduced predator, but that these behavioural shifts do not prevent patterns of natural selection from changing in experimental populations.

Caribbean *Anolis* lizards are ideal subjects for examining the evolutionary role of behaviour. Comparative and experimental studies indicate that populations alter their habitat use in response to the presence of competing or predatory species<sup>12–14</sup>; observations reveal that individuals change their behaviour over short periods of

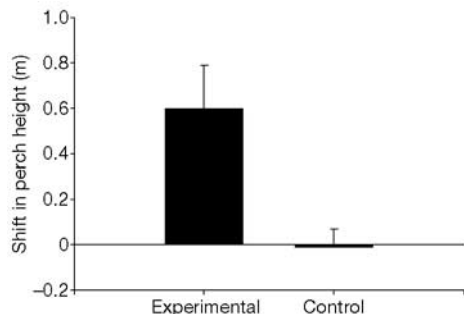


time depending on the local presence of other species (reviewed in ref. 15). Previous studies on *Anolis sagrei*, a small lizard species that usually occurs on or near the ground, showed that it moved higher in the vegetation in the presence of the larger ground-dwelling curly-tailed lizard, *Leiocephalus carinatus*<sup>16,17</sup>. Because *L. carinatus* eats *A. sagrei*<sup>18</sup>, we predict that the presence of *L. carinatus* may lead to shifts in the pattern of natural selection in experimental populations. In particular, because body size and relative limb length are correlated with sprinting ability in anoles and other lizards<sup>19</sup>, we predict that the presence of *L. carinatus* may favour larger and longer-legged *A. sagrei* individuals that are better able to escape. In addition, gape-limitation of *L. carinatus*<sup>18</sup> would also favour larger *A. sagrei* individuals. On the other hand, by moving into the vegetation, *A. sagrei* may avoid predation by the more terrestrial *L. carinatus* and thus preclude selection on these traits. Previous studies comparing populations in the presence or absence of *L. carinatus* have detected differences in population size and trait distributions<sup>16,17</sup> consistent with the hypothesis of predation-driven natural selection, but selection on individual traits has never previously been investigated.

To test these hypotheses, in June 2003 we staged an introduction of curly-tailed lizards to six small islands in the Bahamas. Six other islands served as controls. Immediately before the introductions, *A. sagrei* were captured, measured and marked on each island. Before the introductions, nearly half of the *A. sagrei* observed were on the ground and experimental and control islands did not differ (Fig. 1;  $F_{1,10} = 0.39$ ,  $P = 0.54$ ).



**Figure 1** Differences in proportion of animals observed on the ground in experimental and control populations. Values are the mean and one standard error of mean values for each island. Data, from 12 islands in May and November and 4 islands in July, are for all individuals >33 mm (the size cut-off for marking individuals) in May and all marked individuals in July and November.

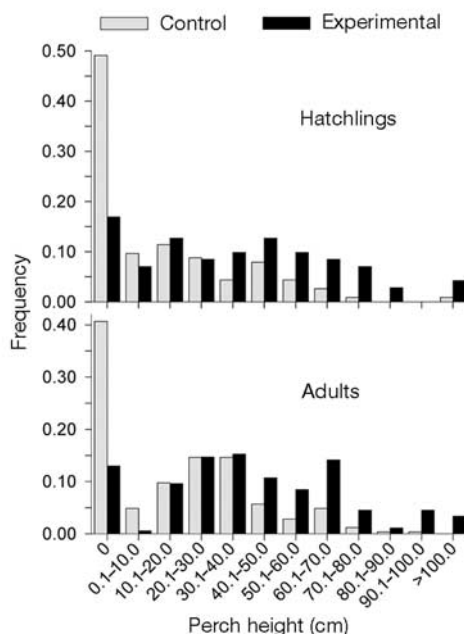


**Figure 2** Change in perch height ( $\pm 1$  s.e.) 10 min after introduction of either a live predatory lizard (experimental) or an inanimate object of the same size (control). Values are means from all individuals across four islands; statistical analyses nested individuals within islands.

When we initially introduced *L. carinatus* to the experimental islands, we released each individual predator 0.5–1.0 m in front of an adult *A. sagrei* in 10-min trials. Although *L. carinatus* almost certainly has not occurred on these islands during the lifetime of any then alive individuals, *A. sagrei* nonetheless responded to *L. carinatus* by moving higher in the vegetation (Fig. 2;  $F_{4,39} = 2.27$ ,  $P = 0.040$ ). Moreover, *A. sagrei* never moved towards *L. carinatus*, whereas on control islands, individuals frequently moved towards the control object (mean number of movements towards introduced object on control islands,  $\bar{x} (\pm 1 \text{ s.e.}) = 1.86 \pm 0.51$ ;  $F_{4,37} = 19.71$ ,  $P < 0.001$ ).

This initial reaction was not permanent, as *A. sagrei* continued to use the ground, presumably when not perceiving a threat from *L. carinatus*. Over time, however, *A. sagrei* became increasingly arboreal on the experimental islands. Six weeks after the introduction, use of the ground on experimental islands was reduced by a third relative to control islands; by six months, only 12% of lizards were observed on the ground on experimental islands versus 34% on controls (Fig. 1), a difference that is statistically significant ( $F_{1,11} = 7.82$ ,  $P = 0.010$ ). Trends for perch height were the same: no difference existed between experimental and control islands in May ( $F_{1,10} = 1.39$ ,  $P = 0.27$ ), but by November, *A. sagrei* on experimental islands were perching much higher than on controls (Fig. 3; mean perch height 0.20 versus 0.10 m;  $F_{1,10} = 36.56$ ,  $P < 0.0001$ ), a difference that is also seen in hatchling lizards (Fig. 3;  $F_{1,9} = 6.45$ ,  $P = 0.016$  (no hatchlings were seen on one island)).

On each island, we measured selection gradients<sup>20,21</sup> on body size and relative limb length (the latter was measured only on males because gravidity makes this an unreliable measurement for females). As predicted, experimental and control islands differed in patterns of selection on both male hindlimb length ( $F_{1,6} = 9.05$ ,  $P = 0.012$ ) and female body size ( $F_{1,8} = 3.72$ ,  $P = 0.045$ ). These differences resulted because on all experimental islands, selection favoured (that is, positive selection gradient) larger females and males with longer legs (Fig. 4). By contrast, no consistent pattern



**Figure 3** Distribution of perch heights on experimental and control populations six months after the introduction of predatory curly-tailed lizards to experimental islands. Distributions are for all individuals on all islands (Fig. 1 presents means of proportions across islands). 'Adults' refers to animals  $\geq 36.5$  mm svl, whereas hatchlings are individuals  $\leq 30$  mm.

was apparent for male body size on experimental islands, nor for any of the traits on control islands.

Two lines of evidence indicate that differences in perch height between experimental and control islands result from individual behavioural shifts rather than selection against lizards that perch at inappropriate heights on islands with curly-tailed lizards. Four islands—two experimental and two control—were intensively studied with respect to the perching characteristics of their *A. sagrei* individuals. First, among individuals surviving to November, the shift to higher perches from the pre-introduction state was greater on experimental than on control islands ( $F_{1,2} = 36.89$ ,  $P = 0.013$ ). Second, among these four islands, there was no evidence that selection favoured individuals that initially perched higher to a greater extent on experimental islands relative to controls: the highest and lowest selection gradients for initial perch height were on the two experimental islands, with the control islands being intermediate.

Our results constitute, to the best of our knowledge, the first experimental test of the hypothesis that behavioural shifts in new environments forestall natural selection. We find little support for this hypothesis. *Anolis sagrei* exhibits a marked behavioural shift in habitat use in the presence of the predatory *L. carinatus*, a shift that is as strong in hatchling lizards as it is in adults. Because *A. sagrei* on experimental islands still used the ground, albeit to an increasingly lesser extent, these habitat shifts were not enough to prevent substantial mortality of *A. sagrei* on experimental islands (mean number of marked individuals that failed to survive ( $\bar{x}$ ) = 0.66, range: 0.47–0.94; by contrast, mortality was variable and sometimes quite low on control islands,  $\bar{x} = 0.45$ , range = 0.09–0.79). This high rate of mortality provided the opportunity for selection to operate, and indeed it did in the direction predicted: the presence of the predatory *L. carinatus* favoured longer-legged male lizards, which can run faster, and favoured larger females, which are both

faster and harder to subdue and ingest. Although no selection on size in males was detected, this was not entirely unexpected: larger males may be more difficult to subdue and ingest, but this advantage is probably countered by the greater vulnerability of larger males resulting from their conspicuous territorial behaviour<sup>22</sup>.

Although the habitat shift exhibited by *A. sagrei* did not prevent selective mortality, it does set the stage for future, behaviour-driven evolutionary change. In contrast to its potentially constraining role, behavioural change also may promote evolutionary change—by changing the way organisms interact with the environment, behaviour may alter selective pressures<sup>1–3</sup>. Such a pattern eventually may apply on our islands as *A. sagrei* becomes increasingly arboreal, a process that previous work suggests will continue for at least several years<sup>17</sup>. By shifting higher into the vegetation, *A. sagrei* uses narrower perches<sup>14,17</sup>. Comparisons of species diverging for millions of years and of populations diverging for thousands of years reveal a consistent pattern: anoles adapt to using narrower surfaces by evolving shorter limbs<sup>15,23</sup>. Functional studies illuminate the underlying cause: on broad surfaces, long limbs provide maximal sprinting and jumping capabilities, whereas on narrow surfaces, short limbs permit agile movement on narrow and irregular surfaces<sup>24</sup>. Therefore, we predict that in future generations, the pattern of selective differences that we have observed will reverse; by moving higher into the vegetation, *A. sagrei* on experimental islands may almost completely escape predation by curly-tailed lizards, but will be forced to use narrower substrates that will favour individuals with shorter, rather than longer, limbs.

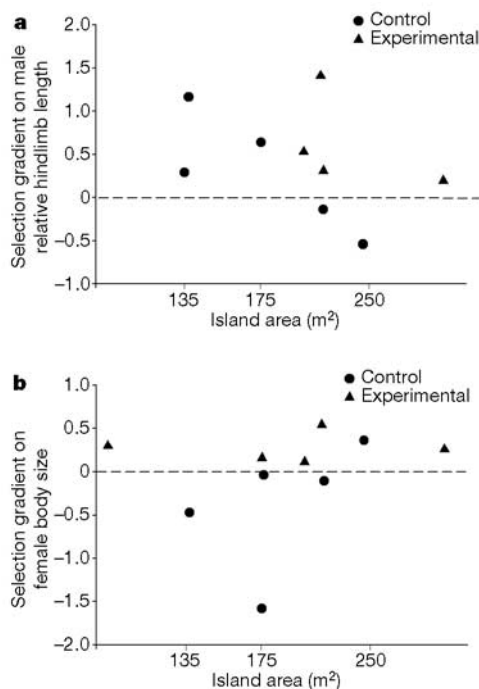
The role of behaviour in evolution is thus potentially not simple. Organisms may often alter their behaviour to avoid new selective agents such as an introduced predator, but such shifts can only forestall selection if they succeed in removing individuals from the new selective agent quickly enough. In our study, it is certainly possible that in the absence of the behavioural shift in habitat, mortality and perhaps selection would have been higher. Nonetheless, even with the habitat shift, mortality rates in populations exposed to the predator were high and provided substantial opportunity for selection to operate. As this behavioural shift continues to habitats increasingly inaccessible to the predator, selection may well change direction: even when behaviour fails to preclude selection of one kind in the short run, it can lead to another kind of selection in the long run. □

## Methods

We located 12 small islands (vegetated area: 104–324 m<sup>2</sup>) lacking curly-tailed lizards in creeks near Snake Cay and Buckaroon Bay, Abaco, Bahamas. Curly-tailed lizards occur on larger islands within these creeks and occasionally colonize islands of the size studied. To ensure that the islands in the two treatments exhibited a similar range of characteristics<sup>25</sup>, we paired islands mainly on the basis of vegetated area, while also considering *A. sagrei* population size, per cent vegetation cover and location (Snake Cay versus Buckaroon Bay), and randomly selected one island from each pair to receive curly-tailed lizards. Adult curly-tailed lizards were collected from nearby areas and introduced on 1–7 June 2003, in proportion to the number of *A. sagrei* on each island (which is highly correlated with island area,  $r = 0.77$ ), which was estimated using mark-recapture methods<sup>13</sup> (estimated prereproductive-season population sizes: 22–103; approximately one curly-tailed lizard introduced per seven *A. sagrei*).

On four of the experimental islands, we conducted focal animal observations on individual *A. sagrei* to investigate their immediate reaction to the introduction of curly-tailed lizards. Lizards were approached and an experimental object—either a live curly-tailed lizard ( $n = 24$ ) or, as a control, an inanimate object of approximately the same size ( $n = 23$ )—was placed 0.5–1.0 m from the lizard on the ground and clearly in its visual field. Trials in which the lizard immediately ran away as the object was being delivered were eliminated, as were those in which the introduced curly-tailed lizard immediately fled. Experimental and control trials were alternated, and we tried to arrange trials in pairs in which the second trial involved an *A. sagrei* of the same sex as the previous one and in a similar initial position. Whether the first trial was experimental or control was randomly decided. Lizards were watched for 10 min after introduction of the object. We recorded the change in height off the ground from the beginning of the trial to the end and the number of movements the lizard made in the direction of the introduced object (this measurement was not recorded in several of the early trials). Data were square root transformed before analysis to equalize variances and were analysed with a two-way nested analysis of variance with individuals nested within islands.

Immediately before and six months after the introductions (22–30 November 2003),



**Figure 4** Selection gradients for male hindlimb length (a) and female snout-vent length (b). Gradients could not be calculated on all islands because on some islands all marked members of one sex either survived or died. The covariate island area was included in the statistical model for male hindlimb length (it is negative), but not for female body size (see Methods).

perch height was recorded for all observed individuals on all 12 islands; in addition, perch height was recorded for marked individuals on four islands (two experimental, two control) in July. Islands were visited multiple times but (with a few exceptions) only once per day. Data were taken for the location at which lizards were first observed; lizards that apparently were disturbed (that is, that appeared to be moving in response to our presence) when first seen were not included.

Also before the introductions, we captured lizards (>33 mm snout-vent length (svl)) on each island ( $n = 9-51$ , 342 total). Hindlimb span (length from insertion of limb in body wall to tip of claw on the fourth toe) was measured on males. Lizards were individually marked by injecting elastomer tags (Northwest Marine Technologies) subdermally in two limb segments. In November, we exhaustively sampled individuals on each island to recapture marked individuals. We estimated survival proportion as the fraction of originally marked individuals (that is, marked in May) found during the November study period, divided by the marginal recapture rate for the last complete census (see Supplementary Information for details on calculation of this rate); if this estimate was smaller than the final number actually found, we used the latter instead.

To measure the effect of curly-tailed lizards on survival selection, we treated the six-month interval as an episode of selection and calculated standardized selection gradients<sup>20,21,26</sup> separately for each sex on each island. These coefficients were then used as the data points for subsequent statistical analyses<sup>27</sup>. We also attempted to calculate selection gradient coefficients using logistic regression<sup>26</sup>, but the regression analyses failed to converge for analyses on several islands, so that estimates could not be obtained. For females, gradients were calculated only for svl, whereas for males they were calculated (in a multiple regression) for svl and relative hindlimb length (= residual of hindlimb length versus svl using the regression for individuals from all islands (analysis of covariance detects no heterogeneity of slopes among islands in the relationship between hindlimb length and svl)). Individuals not recaptured were considered to have died, with the exception that because of the loss of one of the two marks, the identity of five surviving females could not be established. On the basis of the frequency of mark loss, we estimate that one to two individuals in the study may have lost both marks and thus may have been incorrectly categorized as non-survivors. We measured selection at the intermediate time of six months, when mean survival on the islands averaged 45%, because we expected that this period was long enough for selective differences to become apparent, but not so long that marked cohorts would have mostly or entirely disappeared, thereby vitiating measurement of selection.

To test statistically the effect of *L. carinatus* on *A. sagrei* traits, we analysed selection gradient coefficients and mean perch heights from each island (one value per island) with the null hypothesis that mean values would not differ between experimental and control populations<sup>28,29</sup>. We first ran analyses of covariance on each dependent variable using island area (log transformed) as a covariate. This covariate was deleted from the model when  $P > 0.10$ ; it was retained only in the model for male relative hindlimb length. Because the variance of an estimated regression coefficient is inversely proportional to sample size, each selection coefficient for a given island was weighted by the number of individuals measured on that island; the weighting method does not change the degrees of freedom in the analysis<sup>30</sup> and statistical significance at the 0.05 level is the same in unweighted analyses. All *P*-values are one-tailed based on a priori hypotheses on the direction of the effect of *L. carinatus* on a given variable.

Received 21 July; accepted 16 September 2004; doi:10.1038/nature03039.

- Wyles, J. S., Kunkel, J. G. & Wilson, A. C. Birds, behavior, and anatomical evolution. *Proc. Natl Acad. Sci. USA* **80**, 4394–4397 (1983).
- Wcislo, W. T. Behavioral environments and evolutionary change. *Annu. Rev. Ecol. Syst.* **20**, 137–169 (1989).
- West-Eberhard, M. J. Phenotypic plasticity and the origins of diversity. *Annu. Rev. Ecol. Syst.* **20**, 249–278 (1989).
- Wake, D. B., Roth, G. & Wake, M. H. On the problem of stasis in organismal evolution. *J. Theor. Biol.* **101**, 211–224 (1983).
- Brandon, R. in *The Role of Behavior in Evolution* (ed. Plotkin, H. C.) 51–71 (MIT Press, Cambridge, USA, 1988).
- Huey, R. B., Hertz, P. E. & Sinervo, B. Behavioral drive versus behavioral inertia in evolution: A null model approach. *Am. Nat.* **161**, 357–366 (2003).
- Eldredge, N. *Macroevolutionary Dynamics: Species, Niches, & Adaptive Peaks* (McGraw-Hill, New York, 1989).
- Coope, G. R. in *Extinction Rates* (eds Lawton, J. H. & May, R. M.) 55–74 (Oxford Univ. Press, Oxford, 1995).
- Leviton, J. S. *Genetics, Paleontology, and Macroevolution* 2nd edn (Cambridge Univ. Press, Cambridge, UK, 2001).
- Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
- Webb, C. O., Ackerly, D. D., McPeck, M. A. & Donoghue, M. J. Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.* **33**, 475–505 (2003).
- Schoener, T. W. Presence and absence of habitat shift in some widespread lizard species. *Ecol. Monogr.* **45**, 233–258 (1975).
- Palcala, S. W. & Roughgarden, J. Population experiments with the *Anolis* lizards of St. Maarten and St. Eustatius. *Ecology* **66**, 129–141 (1985).
- Losos, J. B. & Spiller, D. Differential colonization success and asymmetrical interactions between two lizard species. *Ecology* **80**, 252–258 (1999).
- Losos, J. B. Integrative approaches to evolutionary ecology: *Anolis* lizards as model systems. *Annu. Rev. Ecol. Syst.* **25**, 467–493 (1994).
- Schoener, T. W., Spiller, D. A. & Losos, J. B. Predators increase the risk of catastrophic extinction of prey populations. *Nature* **412**, 183–186 (2001).
- Schoener, T. W., Spiller, D. A. & Losos, J. B. Predation on a common *Anolis* lizard: Can the food-web effects of a devastating predator be reversed? *Ecol. Monogr.* **72**, 383–408 (2002).

- Schoener, T. W., Slade, J. B. & Stinson, C. H. Diet and sexual dimorphism in the very catholic lizard genus *Leiocephalus* of the Bahamas. *Oecologia* **53**, 160–169 (1982).
- Garland, T. Jr & Losos, J. B. in *Ecological Morphology: Integrative Organismal Biology* (eds Wainwright, P. C. & Reilly, S.) 240–302 (Univ. Chicago Press, Chicago, USA, 1994).
- Lande, R. & Arnold, S. J. The measurement of selection on correlated characters. *Evolution* **37**, 1210–1226 (1983).
- Brodie, E. D. III, Moore, A. J. & Janzen, F. J. Visualizing and quantifying natural selection. *Trends Ecol. Evol.* **10**, 313–318 (1995).
- Schoener, T. W. & Schoener, A. The ecological correlates of survival in some Bahamian *Anolis* lizards. *Oikos* **39**, 1–16 (1982).
- Williams, E. E. in *Lizard Ecology: Studies of a Model Organism* (eds Huey, R. B., Pianka, E. R. & Schoener, T. W.) 326–370 (Harvard Univ. Press, Cambridge, USA, 1983).
- Irschick, D. J. & Losos, J. B. Do lizards avoid habitats in which performance is submaximal? The relationship between sprinting capabilities and structural habitat use in Caribbean anoles. *Am. Nat.* **154**, 293–305 (1999).
- Hurlbert, S. H. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**, 187–211 (1984).
- Janzen, F. J. & Stern, H. S. Logistic regression for empirical studies of multivariate selection. *Evolution* **52**, 1564–1571 (1998).
- DeWitt, T. J. & Langerhans, R. B. Multiple prey traits, multiple predators: keys to understanding complex community dynamics. *J. Sea Res.* **49**, 143–155 (2003).
- Rundle, H. D., Vamosi, S. M. & Schluter, D. Experimental test of predation's effect on divergent selection during character displacement in sticklebacks. *Proc. Natl Acad. Sci. USA* **100**, 14943–14948 (2003).
- Bolnick, D. I. Can intraspecific competition drive disruptive selection? An experimental test in natural populations of sticklebacks. *Evolution* **58**, 608–618 (2004).
- Littell, R. C., Freund, R. J. & Spector, P. C. *SAS System for Linear Models* 3rd edn (SAS Institute Inc., Cary, North Carolina, 1999).

Supplementary Information accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank the National Science Foundation and the National Geographic Society for support, J. Chase, T. Knight and B. Pinder for assistance, R. B. Langerhans for suggesting the approach to study selection and helping in its implementation, D. Bolnick, J. Chase, B. Fitzpatrick, F. Janzen, T. Knight, B. Langerhans, M. Leal, M. Turelli and the Turelli labgroup, for constructive comments on previous drafts, and the Bahamian government for permission to conduct this research.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to J.B.L. (losos@biology.wustl.edu).

## Magnetoreception and its trigeminal mediation in the homing pigeon

Cordula V. Mora<sup>1\*</sup>, Michael Davison<sup>2</sup>, J. Martin Wild<sup>3</sup> & Michael M. Walker<sup>1</sup>

<sup>1</sup>School of Biological Sciences, <sup>2</sup>Department of Psychology, <sup>3</sup>Anatomy Department, School of Medicine, University of Auckland, Private Bag 92019, Auckland, New Zealand

\* Present address: Department of Biology, Coker Hall, CB 3280, University of North Carolina, Chapel Hill, North Carolina 27599-3280, USA

Two conflicting hypotheses compete to explain how a homing pigeon can return to its loft over great distances. One proposes the use of atmospheric odours<sup>1</sup> and the other the Earth's magnetic field<sup>2–4</sup> in the 'map' step of the 'map and compass' hypothesis of pigeon homing<sup>5</sup>. Although magnetic effects on pigeon orientation<sup>6,7</sup> provide indirect evidence for a magnetic 'map', numerous conditioning experiments<sup>8</sup> have failed to demonstrate reproducible responses to magnetic fields by pigeons. This has led to suggestions that homing pigeons and other birds have no useful sensitivity to the Earth's magnetic field<sup>9–11</sup>. Here we demonstrate that homing pigeons (*Columba livia*) can discriminate between the presence and absence of a magnetic anomaly in a conditioned choice experiment. This discrimination is impaired by attachment of a magnet to the cere, local anaesthesia of the upper beak area, and bilateral section of the ophthalmic branch of

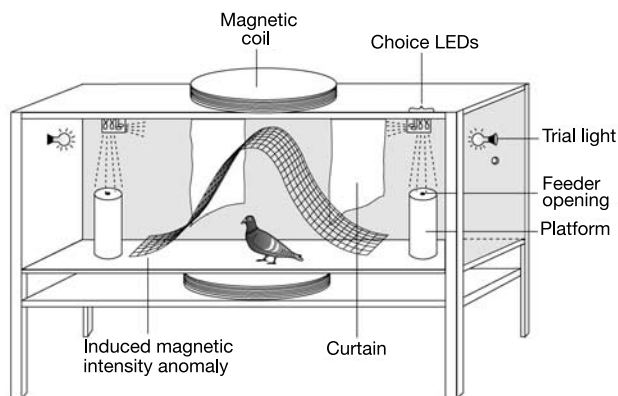


the trigeminal nerve, but not of the olfactory nerve. These results suggest that magnetoreception (probably magnetite-based) occurs in the upper beak area of the pigeon. Traditional methods of rendering pigeons anosmic might therefore cause simultaneous impairment of magnetoreception so that future orientation experiments will require independent evaluation of the pigeon's magnetic and olfactory systems.

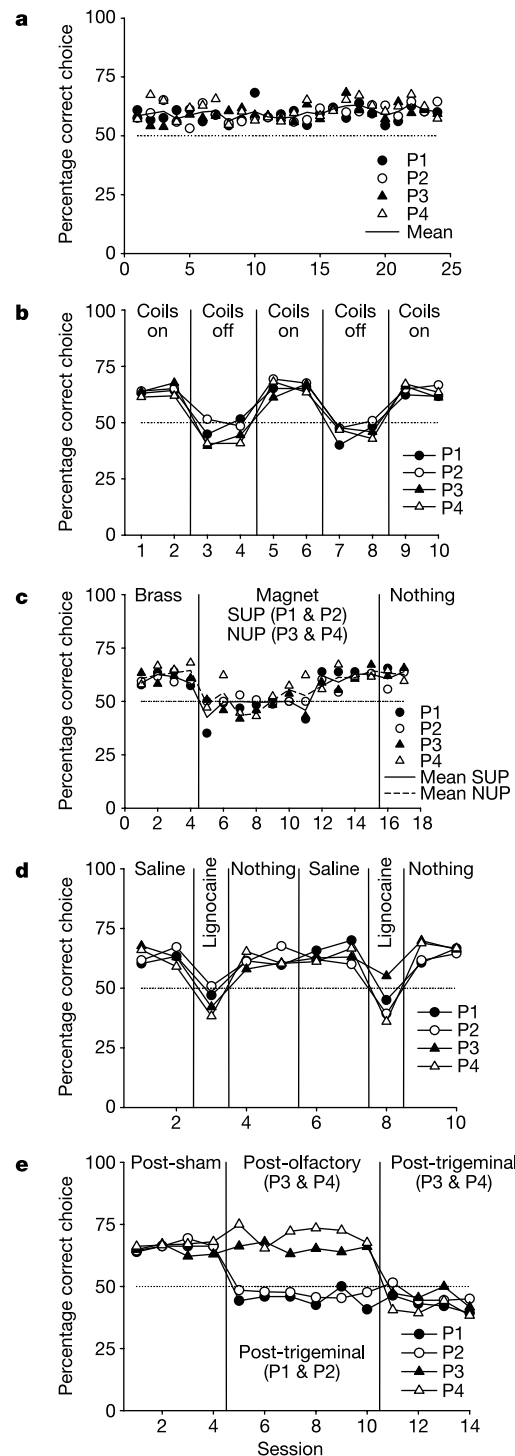
We made a series of modifications to an existing operant conditioning procedure<sup>12</sup> to fulfil two conditions that seem to be vital for magnetic discrimination learning in non-avian species<sup>13–15</sup>. These are that (1) the magnetic stimulus discriminated is a localized, non-uniform magnetic anomaly superimposed on the uniform background field of the Earth, and (2) movement by the experimental subjects is necessary to produce the behavioural response measured in the experiments. Although this combination of experimental parameters mitigates against rapid achievement of powerful discrimination by separating the stimulus, response and reinforcement in both space and time—compared with standard key-pecking experiments—failure to fulfil either or both of the above conditions has characterized all the unsuccessful or irreproducible attempts to condition pigeons and many other species to magnetic fields<sup>8,12</sup>.

Using a Yes–No signal-detection procedure<sup>16</sup> (see Supplementary Methods), four individually trained pigeons were required to discriminate between the presence and absence of an induced magnetic field anomaly while freely walking in a wooden tunnel (Fig. 1). The intensity profile of the anomaly was 'wave-shaped' and peaked in the centre of the tunnel at 189 micro tesla ( $\mu\text{T}$ ) (background level of  $44\ \mu\text{T}$ ) with an inclination of  $-80^\circ$  (background level of  $-64^\circ$ ). The birds were conditioned to jump onto a platform at one end of the tunnel when the anomaly was present and onto an identical platform at the other end of the tunnel when the anomaly was absent. Choice of the correct platform was rewarded with food whereas incorrect choices were punished with a time penalty.

Over 24 consecutive conditioning sessions, the percentage of correct choices made by the pigeons fell mostly between 55% and 65% (mean of 59.81% with upper and lower 95% confidence limits 58.78% and 60.83%) and occasionally approached 70% (Fig. 2a). Mean discrimination performance was therefore significantly greater than the chance level of 50% ( $P < 0.0001$ ) with the birds



**Figure 1** Experimental tunnel used in conditioned choice discrimination of magnetic stimuli. Four individually trained homing pigeons (*C. livia*) discriminated the presence and absence of a magnetic field anomaly, which was wave-shaped in its intensity profile and located centrally in the tunnel (peak intensity and inclination varied from  $44\ \mu\text{T}$  to  $189\ \mu\text{T}$  and  $-64^\circ$  to  $-80^\circ$  respectively). The birds were required to mount one of the two feeder platforms depending on the magnetic field present in the tunnel during a discrete trial. Correct choices were rewarded with food from a feeder opening in the top of the platform whereas incorrect choices resulted in a time penalty.



**Figure 2** Percentage of correct choices by four individually trained homing pigeons (P1–P4) discriminating the presence and absence of a magnetic field anomaly over consecutive daily sessions. The horizontal line at 50% indicates the level of chance performance. **a**, Baseline magnetic choice discrimination performance. **b**, Coil control sessions with resistors replacing coil load (to test for use of extraneous cues in discrimination). **c**, Magnetic impairment sessions in which brass weights (control) or magnets (SUP for P1 and P2, NUP for P3 and P4) have been attached to the top of the cere. **d**, Anaesthetic impairment sessions in which the olfactory mucosa is bathed in a 2% lignocaine hydrochloride solution, physiological saline (control) or nothing. **e**, Nerve sectioning sessions in which sham operations followed by conditioning sessions act as controls for subsequent bilateral sectioning of the olfactory nerve (NI) and the ophthalmic branch (V1) of the trigeminal nerve (NV).

clearly being able to discriminate the presence and absence of the magnetic anomaly in the tunnel. Mean discrimination performance continued to improve ( $t_{\text{obs}} = 9.112$ ,  $P < 0.001$ ) over the control sessions of the subsequent impairment experiments (Fig. 2b–e), reaching 75% after olfactory nerve sectioning (Fig. 2e). This result is close to the 80% value reported as the upper range of discrimination performance seen in pigeons for difficult discrimination tasks with other sensory systems<sup>17</sup>. Acquisition of discrimination is not evident in Fig. 2a because the birds had gained extended exposure to the stimuli and their associated reinforcements while the procedure was being developed. As a consequence, their percentage of correct choices abruptly rose above 50% once the final modification to the discrimination procedure was made that revealed discrimination of the stimuli by the birds.

To control for the possibility that the birds might use cues from the apparatus that were unrelated to the magnetic field anomaly, the birds were tested after substitution of equivalent resistors for the magnetic coils in the electrical circuit (Fig. 2b). Over ten sessions, the birds readily discriminated the presence and absence of the anomaly when the coils were carrying current from the power supply (sessions 1, 2, 5, 6, 9, 10). In contrast, they failed to make correct responses more than 50% of the time when the resistors were present (sessions 3, 4, 7, 8;  $F_{\text{obs}} = 266.21$ ,  $P < 0.0001$ ). Successful impairment of the pigeons' magnetic discrimination performance (see below) also argues against the use of extraneous cues associated with the experimental apparatus.

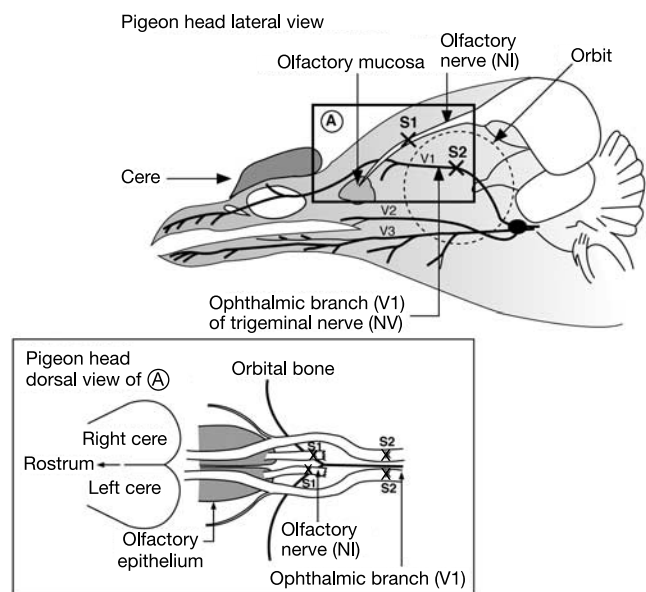
Successful discrimination of the anomaly by the birds permitted a series of impairment experiments to investigate the probable location and mechanism of the pigeons' magnetic sense. First, small but strong rare-earth magnets (NdFeB cylinders of 3-mm diameter and 2-mm length with a field strength of 2,500  $\mu\text{T}$  at 1-cm distance) were attached to the cere for the duration of each of eleven conditioning sessions (sessions 5–15; Fig. 2c). The orientation of the magnet was south-up (SUP) for two pigeons and north-up (NUP) for the other two birds. With this magnet attached to the cere, the pigeons' discrimination performance immediately dropped to chance level ( $P = 0.1103$ ), indicating that the magnet impaired the ability of the pigeons to discriminate magnetic fields. Control sessions where the birds carried brass weights of equivalent size (sessions 1–4; Fig. 2c) or, following magnetic impairment, nothing at all (sessions 16–17; Fig. 2c) showed an average discrimination performance significantly greater than chance (both  $P < 0.0001$ ) and indistinguishable from baseline performance (Fig. 2a). Mean performance during the brass weight sessions was significantly better than that with attached magnets ( $F_{\text{obs}} = 8.00$ ,  $P = 0.0048$ ). Magnetic impairment was temporary, however, as performance gradually recovered over several sessions (sessions 9–11; Fig. 2c) to the level achieved previously. This change in behaviour over time was confirmed by a large, statistically reliable covariance (36.94,  $Z_{\text{obs}} = 1.88$ ,  $P = 0.0297$ ) for the mean performance over the magnet sessions, whereas the covariances over the brass weight and untreated sessions were effectively zero. No effect of the orientation of the magnet (NUP or SUP) was detected ( $F_{\text{obs}} = 0.70$ ,  $P = 0.4071$ ).

Second, to assess whether observed effects of the magnet were on hypothesized magnetoreceptors located either in the eye<sup>18</sup> or in the upper beak area<sup>19,20</sup>, we locally anaesthetized the olfactory cavity using a 2% lignocaine hydrochloride solution. The birds performed normally after bathing the olfactory cavity in physiological saline (sessions 1, 2, 6, 7; Fig. 2d) and during untreated control sessions (sessions 4, 5, 9, 10; Fig. 2d), but performed poorly after local anaesthesia of the olfactory cavity (sessions 3, 8,  $F_{\text{obs}} = 28.93$ ,  $P < 0.0001$ ; Fig. 2d). Discrimination was thus reversibly abolished by anaesthesia of the olfactory cavity.

The final experiment aimed to identify the afferent nerve that carries magnetic field information to the brain by separately sectioning the trigeminal and olfactory nerves before they become

closely associated in the upper beak area (Fig. 3). Sham operations followed by four conditioning sessions (sessions 1–4; Fig. 2e) in which the birds continued to discriminate the magnetic field anomaly after recovery (both groups  $P < 0.0001$ ) provided a control for the effects of surgery on discrimination. Bilateral sectioning of the ophthalmic branch (V1) of the trigeminal nerve (NV) in two pigeons was followed by discrimination performance slightly below chance level (sessions 5–14,  $P = 0.0034$ ; Fig. 2e), whereas bilateral sectioning of the olfactory nerve (NI) in the other two birds had no such effect (sessions 5–10,  $P < 0.0001$ ; Fig. 2e). Discrimination subsequently failed in the olfactory nerve-sectioned birds after bilateral section of the ophthalmic branch of the trigeminal nerve (sessions 11–14,  $P = 0.0029$ ; Fig. 2e). The differences in discrimination performance between these three treatment stages, as designated by the vertical lines in Fig. 2e, were significant ( $F_{\text{obs}} = 222.32$ ,  $P < 0.0001$ ).

The results from these impairment experiments are consistent with the detection of magnetic fields using magnetite located in the front of the head<sup>19–21</sup> and electrophysiological experiments that have demonstrated responses to magnetic field stimuli in V1 of the bobolink<sup>22</sup>. They are not consistent, however, with the use of light-mediated magnetoreception for magnetic discrimination. We therefore suggest that magnetoreception in pigeons, and possibly in birds in general, is at least in part dependent on magnetite located in the upper beak area with the ophthalmic branch of the trigeminal nerve carrying magnetic field information to the brain. The magnetic sense of pigeons thus shares a similar location (the nasal region) and afferent nerve (the trigeminal) with at least the teleost fishes<sup>15,23</sup>. Field experiments<sup>8,24</sup> and a theoretical model<sup>25</sup> suggest that this magnetite-based system could form the basis for a magnetic 'map' in homing pigeons. Now that the pigeon's magnetic sense can finally be studied in the laboratory in the same way as other sensory modalities, the sensitivity of homing pigeons to magnetic fields can be analysed in psychophysical studies and compared with theoretical requirements for use in a 'map'. The transduction mechanism of the pigeon's magnetic



**Figure 3** Lateral and dorsal (insert) views of the anatomical relationship between the ophthalmic branch (V1) of the trigeminal nerve (NV), the olfactory nerve (NI) and the olfactory mucosa in the head region of the homing pigeon (*C. livia*). Crosses S1 and S2 indicate the points of bilateral nerve sectioning for the olfactory and ophthalmic nerves, respectively.

compass remains to be determined, and we note that a separate, light-mediated pathway has been suggested from work with migratory birds<sup>8,18</sup>.

Our results also suggest that interpretation of previous magnetic and olfactory impairment experiments in homing pigeons and possibly other species, such as sea turtles<sup>26</sup>, albatrosses<sup>27</sup> and petrels<sup>28,29</sup> requires caution. There is a possibility that attached magnets, even when applied close enough to the magnetoreceptor to elicit an effect on orientation behaviour, will produce only temporary or incomplete impairment of the magnetic sense. Moreover, as first suggested by ref. 30, induction of anosmia by local anaesthesia, application of zinc sulphate to the olfactory mucosa, or olfactory nerve sectioning at the base of the beak, might simultaneously impair the magnetic sense because the olfactory and trigeminal structures are in close proximity in the nasal cavity (Fig. 3). If magnetite is present in the olfactory mucosa of the pigeon, as it is in rainbow trout<sup>15,23</sup>, or if the magnetoreceptors are found in nearby regions<sup>19</sup>, any of the above treatments or even mechanical plugging of the nostrils may potentially directly affect the operation of the magnetoreceptors themselves. To overcome the above difficulties in interpreting magnetic and olfactory impairment experiments and resolve the debate over use of magnetic and olfactory stimuli in homing pigeons, independent sectioning or blocking of the olfactory nerve and the ophthalmic branch of the trigeminal nerve will be required. We suggest that our work provides the basis for detailed studies of both the operation and use of the magnetic sense in homing pigeons and possibly migratory bird species. □

## Methods

### Experimental birds

Two female and two male experienced racing pigeons, aged 3–7 yr, were given water *ad libitum* while their food intake was restricted to maintain each animal at  $85\% \pm 15\%$  of its free-feeding body weight.

### Experimental apparatus

Daily conditioning sessions were conducted in a wooden tunnel<sup>12</sup> (3.30 m length  $\times$  1.06 m width  $\times$  0.96 m height) (Fig. 1) with two identical platforms (48 cm high) located at opposite ends. Each platform was fitted with microswitches to detect the weight of a pigeon, and a rotating feeder disc controlling access to the food reservoir through a feeder opening. Two transparent curtains suspended from the ceiling of the tunnel forced the pigeons to walk, not fly, between the platforms. Trial lights at either end of the tunnel signalled the stimulus-sampling period, while four white light-emitting diodes (LEDs) directly over each platform signalled both the availability of the platforms for the choice response and whether or not the correct response had been made. Insulation of the coils and location of electronic equipment in a separate room minimized the possibility that the birds could detect any thermal or auditory cues.

### Discriminative stimuli

The magnetic field intensity anomaly was produced by two identical single-wrapped (100 turns of 0.05-mm copper wire) circular coils of 1.11-m diameter, located just above and below the centre of the tunnel (Fig. 1). With the coils wired in parallel and a constant, direct current of 3.0 A passing through them, the fields projected into the central region of the tunnel by the two coils summed to produce a distinctive magnetic field anomaly in which both intensity and inclination changed markedly from background values of  $44 \mu\text{T}$  and  $-64^\circ$  to  $189 \mu\text{T}$  and  $-80^\circ$ , respectively. During the control experiment, five  $6.8 \Omega$  resistors were wired in parallel to match the effect of the coils' load (1.5 A at a resistance of  $1.36 \Omega$  per coil) on the power source.

### Nerve sectioning

Under general anaesthesia, the paired olfactory nerves (NI) were approached via a midline burr hole made in the rostral skull. The thin, bony canals housing the nerves were opened and a 2-mm section of each nerve was removed (Fig. 3). The ophthalmic branch (V1) of the trigeminal nerve (NV) behind the left eye was approached after an incision through the upper eyelid and orbital fascia and gentle depression of the globe. A 3-mm piece of nerve was removed (Fig. 3). V1 running behind the right eye was sectioned via the left orbit after removal of a small part of the interorbital bone. V1-sectioning did not impair the ability of the birds to mount the feeder platform or to consume the food reward. Sham operations involved the same procedures, except that the nerves were not cut. All birds were given 60 h to recover from sham surgery and nerve sectioning before conditioned choice testing.

### Statistical analysis

A linear mixed model was fitted to each normally distributed data set using the Statistical

Analysis System (SAS). The model permitted us to test first for differences between the mean percentage of correct choices made by the birds and the 50% correct choice expected from chance performance. We could then examine the data for the occurrence of learning, detected as changes in behaviour over time, and estimate and accommodate any autocorrelation between sessions.

Further data analysis and computer simulations (not shown) were conducted to determine the cause of the slightly below-chance performance in the coils-off sessions in the control experiment, the magnetic and anaesthesia impairment sessions and the sessions after V1 sectioning. These revealed that this below-chance performance resulted from the interaction between the birds switching to a pattern of frequent alternation between platforms independent of stimulus presentation during impairment sessions, and slight, unexpected non-randomness in the balanced quasi-random order of stimulus presentation used in this study.

Received 4 May; accepted 4 October 2004; doi:10.1038/nature03077.

1. Papi, F., Fiore, L., Fiaschi, V. & Benvenuti, S. Olfaction and homing in pigeons. *Monit. Zool. Ital. (N.S.)* **6**, 85–95 (1972).
2. Gould, J. L. The case for magnetic sensitivity in birds and bees (such as it is). *Am. Sci.* **68**, 256–267 (1980).
3. Moore, B. R. Is the homing pigeon's map geomagnetic? *Nature* **285**, 69–70 (1980).
4. Walcott, C. Magnetic orientation in homing pigeons. *IEEE Trans. Magn.* **16**, 1008–1013 (1980).
5. Kramer, G. Wird die Sonnenhöhe bei der Heimfindeorientierung verwertet? *J. Ornithol.* **94**, 201–219 (1953).
6. Walcott, C. in *Animal Migration, Navigation, and Homing* (eds Schmidt-Koenig, K. & Keeton, W. T.) 143–153 (Springer, Heidelberg, 1978).
7. Keeton, W. T., Larkin, T. S. & Windsor, D. M. Normal fluctuations in the earth's magnetic field influence pigeon orientation. *J. Comp. Physiol.* **95**, 95–103 (1974).
8. Wiltschko, R. & Wiltschko, W. *Magnetic Orientation in Animals* (Springer, Heidelberg, 1995).
9. Griffin, D. R. Ecology of migration: is magnetic orientation a reality? *Q. Rev. Biol.* **57**, 293–295 (1982).
10. Wallraff, H. G. Simulated navigation based on assumed gradients of atmospheric trace gases (Models on pigeon homing Part 2). *J. Theor. Biol.* **138**, 511–528 (1989).
11. Papi, F. *Animal Homing* (Chapman and Hall, London, 1992).
12. Carman, G. J., Walker, M. M. & Lee, A. K. Attempts to demonstrate magnetic discrimination by homing pigeons in flight. *Anim. Learn. Behav.* **15**, 124–129 (1987).
13. Walker, M. M. & Bitterman, M. E. Conditioned responding to magnetic fields by honeybees. *J. Comp. Physiol. A* **157**, 67–71 (1985).
14. Walker, M. M. Learned magnetic field discrimination in yellow-fin tuna, *Thunnus albacares*. *J. Comp. Physiol. A* **155**, 673–679 (1984).
15. Walker, M. M. *et al.* Structure and function of the vertebrate magnetic sense. *Nature* **390**, 371–376 (1997).
16. Davison, M. C. & Tustin, R. D. The relation between the generalized matching law and signal-detection theory. *J. Exp. Anal. Behav.* **29**, 331–336 (1978).
17. Delius, J. D. in *Physiology and Behaviour of the Pigeon* (ed. Abs, M.) 327–355 (Academic, London, 1983).
18. Ritz, T., Dommer, D. H. & Phillips, J. B. Shedding light on vertebrate magnetoreception. *Neuron* **34**, 503–506 (2002).
19. Williams, M. N. & Wild, J. M. Trigeminal innervated iron-containing structures in the beak of homing pigeons, and other birds. *Brain Res.* **899**, 243–246 (2001).
20. Fleissner, G. *et al.* Ultrastructural analysis of a putative magnetoreceptor in the beak of homing pigeons. *J. Comp. Neurol.* **458**, 350–360 (2003).
21. Walcott, C., Gould, J. L. & Kirschvink, J. L. Pigeons have magnets. *Science* **205**, 1027–1029 (1979).
22. Beason, R. C. & Semm, P. Does the avian ophthalmic nerve carry magnetic navigational information? *J. Exp. Biol.* **199**, 1241–1244 (1996).
23. Diebel, C. E., Proksch, R., Green, C. R., Neilson, P. & Walker, M. M. Magnetite defines a vertebrate magnetoreceptor. *Nature* **406**, 299–302 (2000).
24. Beason, B. C., Wiltschko, R. & Wiltschko, W. Pigeon homing: effects of magnetic pulses on initial orientation. *Auk* **114**, 405–415 (1997).
25. Walker, M. M. On a wing and a vector: a model for magnetic navigation by homing pigeons. *J. Theor. Biol.* **192**, 341–349 (1998).
26. Papi, F., Luschi, P., Akesson, S., Capogrossi, S. & Hays, G. C. Open-sea migration of magnetically disturbed sea turtles. *J. Exp. Biol.* **203**, 3435–3443 (2000).
27. Mouritsen, H., Huyvaert, K. P., Frost, B. J. & Anderson, D. J. Waved albatrosses can navigate with strong magnets attached to their head. *J. Exp. Biol.* **206**, 4155–4166 (2003).
28. Benhamou, S., Bonadonna, F. & Jouventin, P. Successful homing of magnet-carrying white-chinned petrels released in the open sea. *Anim. Behav.* **65**, 729–734 (2003).
29. Bonadonna, F. & Bretagnolle, V. Smelling home: a good solution for burrow-finding in nocturnal petrels? *J. Exp. Biol.* **205**, 2519–2523 (2002).
30. Schmidt-Koenig, K. in *Physiology and Behaviour of the Pigeon* (ed. Abs, M.) 267–283 (Academic, London, 1983).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** C.V.M. was formerly Cordula V. Haugh. This work was supported by grants from the Auckland University Research Committee. We thank J. Longville for supplying the pigeons, M. Williams for advice on the use of anaesthetics, B. McArdle for help with statistical analysis, V. Ward for graphical support and R. Mora for technical support.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to C.V.M. (cvmora@email.unc.edu).



# Mast cells promote homeostasis by limiting endothelin-1-induced toxicity

Marcus Maurer<sup>1,2,3\*</sup>, Jochen Wedemeyer<sup>1,4,5\*</sup>, Martin Metz<sup>1,2,4</sup>, Adrian M. Piliponsky<sup>4</sup>, Karsten Weller<sup>2</sup>, Devavani Chatterjea<sup>4</sup>, David E. Clouthier<sup>6</sup>, Masashi M. Yanagisawa<sup>7,8,9</sup>, Mindy Tsai<sup>1,4</sup> & Stephen J. Galli<sup>1,4</sup>

<sup>1</sup>Department of Pathology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA

<sup>2</sup>Department of Dermatology, Universität Mainz, D-55101 Mainz, Germany

<sup>3</sup>Department of Dermatology and Allergy, University Hospital Charité, Humboldt University Berlin, D-10117, Germany

<sup>4</sup>Department of Pathology, Stanford University School of Medicine, Stanford, California 94305-5324, USA

<sup>5</sup>Department of Gastroenterology, Hepatology and Endocrinology, Center of Internal Medicine, Medizinische Hochschule Hannover, 30623 Hannover, Germany

<sup>6</sup>Department of Molecular, Cellular and Craniofacial Biology, University of Louisville, Louisville, Kentucky 40292, USA

<sup>7</sup>Department of Molecular Genetics, <sup>8</sup>Howard Hughes Medical Institute, and

<sup>9</sup>Donald W. Reynolds Cardiovascular Clinical Research Center, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390-9050, USA

\* These authors contributed equally to this work

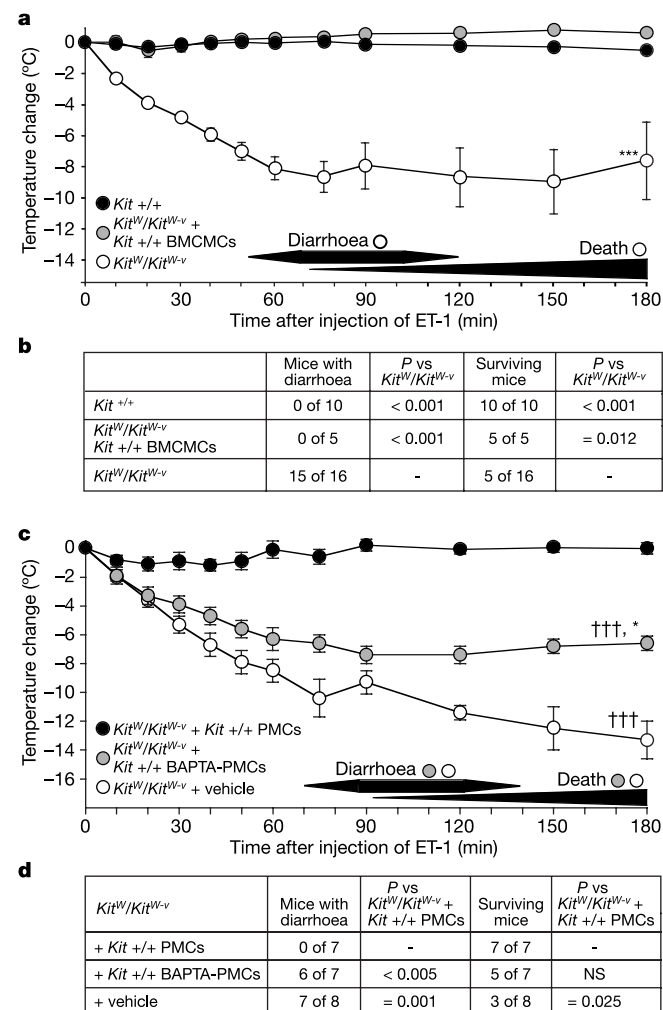
Endothelin-1 (ET-1) is a 21-amino-acid peptide, derived from vascular endothelial cells, with potent vasoconstrictor activity<sup>1</sup>. ET-1 has been implicated in diverse physiological or pathological processes<sup>2,3</sup>, including the vascular changes associated with sepsis<sup>2-5</sup>. However, the factors that regulate ET-1-associated toxicity during bacterial infections, or in other settings, are not fully understood<sup>2-5</sup>. Both the pathology associated with certain allergic and autoimmune disorders<sup>6,7</sup>, and optimal host defence against bacterial and parasitic infections<sup>8-10</sup> are mediated by mast cells. *In vitro*, mast cells can produce ET-1 (ref. 11), undergo ET-1-dependent and endothelin-A receptor (ET<sub>A</sub>)-dependent activation<sup>12,13</sup>, and release proteases that degrade ET-1 (ref. 14). Although the potential relationships between mast cells and the ET-1 system thus may be complex, the importance of interactions between ET-1 and mast cells *in vivo* is obscure. Here we show that ET<sub>A</sub>-dependent mast-cell activation can diminish both ET-1 levels and ET-1-induced pathology *in vivo*, and also can contribute to optimal survival during acute bacterial peritonitis. These findings identify a new biological function for mast cells: promotion of homeostasis by limiting the toxicity associated with an endogenous mediator.

To assess the ability of ET-1 to produce pathology in the presence or absence of mast cells *in vivo*, we administered ET-1 (1.2 nmol in 300 µl of saline) intraperitoneally (i.p.) to mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice and the congenic normal *Kit<sup>+</sup>/Kit<sup>+</sup>* mice<sup>15</sup>. Virtually all of the *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice developed severe hypothermia (a finding also observed during sepsis) and diarrhoea, and 11 of 16 mice died within 3 h after ET-1 injection; by contrast, none of the *Kit<sup>+</sup>/Kit<sup>+</sup>* mice exhibited hypothermia, morbidity or mortality (Fig. 1a, b). Injection of a scrambled peptide instead of ET-1 was without effect in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice, and *Kit<sup>+</sup>/Kit<sup>+</sup>* mice exhibited no morbidity after injection of a tenfold higher dose of ET-1 (that is, 12 nmol) than that which induced severe morbidity in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (Supplementary Fig. 1).

In addition to their profound deficiency in mast cells, *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice are mildly anaemic, lack cutaneous melanocytes, are sterile and virtually lack interstitial cells of Cajal<sup>15-17</sup>. To address the possibility that the differences in ET-1-induced toxicity reflected effects of the *c-kit* mutations in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice other than the

mast-cell deficiency, we also administered ET-1 to a group of *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that had undergone selective repair of their peritoneal mast cell (PMC) deficiency by the i.p. transfer of immature mast cells that were derived *in vitro* (BMCMCs, bone marrow-derived cultured mast cells) from the bone marrow of *Kit<sup>+</sup>/Kit<sup>+</sup>* mice (*Kit<sup>W</sup>/Kit<sup>W-v</sup>* + *Kit<sup>+</sup>/Kit<sup>+</sup>* BMCMC mice)<sup>18,19</sup>. *Kit<sup>W</sup>/Kit<sup>W-v</sup>* + *Kit<sup>+</sup>/Kit<sup>+</sup>* BMCMC mice were completely protected from ET-1-induced morbidity and mortality (Fig. 1a, b). We also found that co-incubation of ET-1 with PMCs *ex vivo* eliminated the ability of the peptide to induce hypothermia and death when administered to *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice *in vivo* (Supplementary Fig. 2).

We next assessed whether degranulation of PMCs is required for the reduction of ET-1-induced toxicity *in vivo* (Fig. 1c, d). *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that had been injected i.p. with peritoneal lavage cells (PLCs) containing  $2.0 \times 10^5$  *Kit<sup>+</sup>/Kit<sup>+</sup>* PMCs per mouse (>90% purity, prepared as in ref. 20) were completely resistant to a subsequent ET-1 injection, whereas vehicle-treated mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice exhibited hypothermia, diarrhoea and mortality. However, when the *Kit<sup>+</sup>/Kit<sup>+</sup>* cells were treated with the



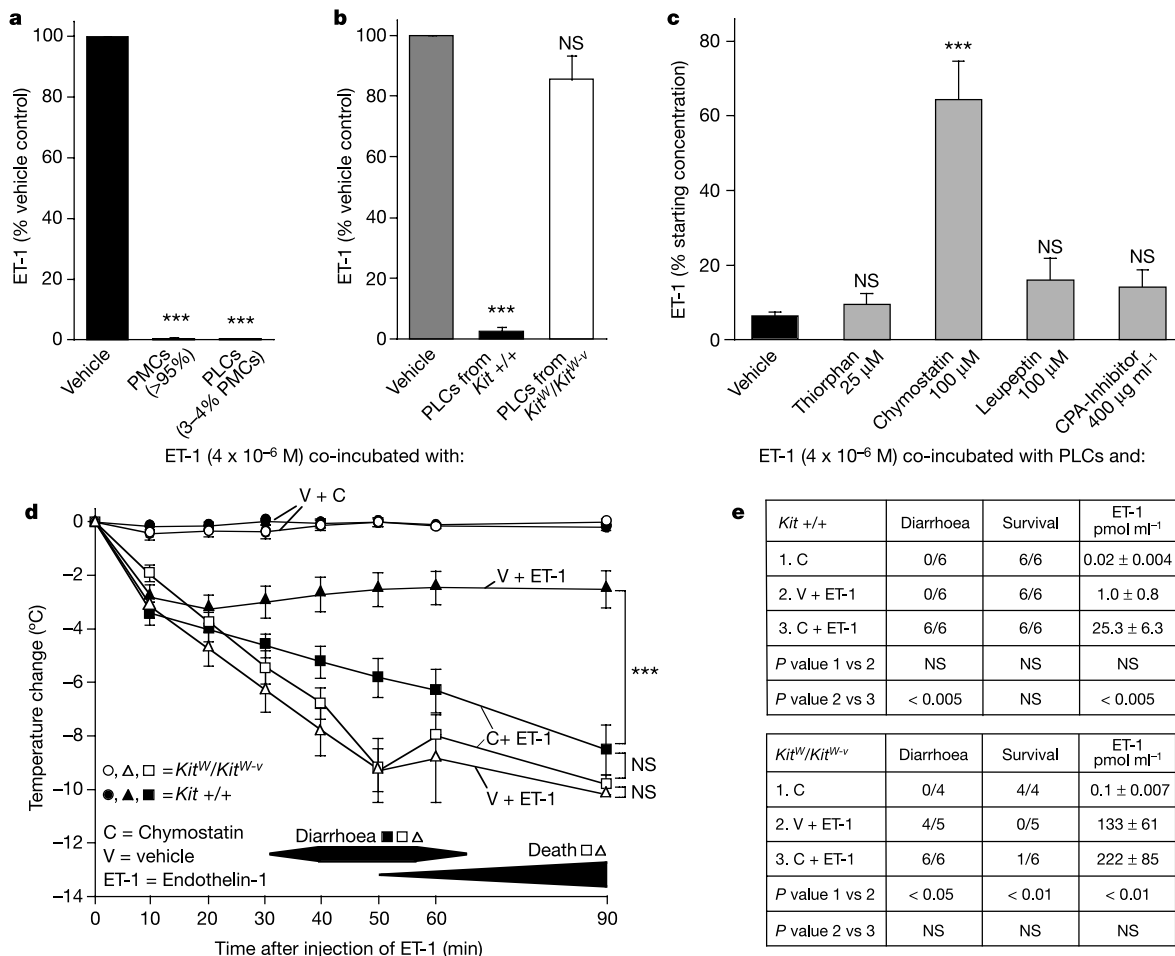
**Figure 1** Mast cells protect mice from morbidity and mortality after the i.p. injection of ET-1. **a–d**, Changes in rectal temperatures (**a**, **c**) and rates of diarrhoea and death (**b**, **d**) after injection of ET-1 (1.2 nmol in 300 µl saline) into wild-type (*Kit<sup>+</sup>/Kit<sup>+</sup>*), mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* and BMCMC-reconstituted *Kit<sup>W</sup>/Kit<sup>W-v</sup>* (*Kit<sup>W</sup>/Kit<sup>W-v</sup>* + *Kit<sup>+</sup>/Kit<sup>+</sup>* BMCMCs) mice (**a**, **b**) or *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (10–12-week-old female) that had received, ~30 min before injection of ET-1, i.p. injections of vehicle or peritoneal cell suspensions (containing 200,000 PMCs per mouse) treated with vehicle (*Kit<sup>W</sup>/Kit<sup>W-v</sup>* + *Kit<sup>+</sup>/Kit<sup>+</sup>* PMCs) or BAPTA-AM (**c**, **d**). \*\*\**P* < 0.001 versus *Kit<sup>+</sup>/Kit<sup>+</sup>* or *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice + *Kit<sup>+</sup>/Kit<sup>+</sup>* + BMCMCs; \**P* < 0.05 versus vehicle; †††*P* < 0.001 versus PMCs. Data in **a**, **b** and **c**, **d** were pooled from three experiments each.

membrane-permeable  $\text{Ca}^{2+}$  chelator BAPTA-AM<sup>12</sup> (50  $\mu\text{M}$  for 30 min) before i.p. injection into *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice, the protective effect of the adoptively transferred PMCs was greatly reduced. BAPTA-AM can inhibit degranulation and histamine release from mouse PMCs stimulated with ET-1 (ref. 12). In confirmation of those results, we found that aliquots of the BAPTA-AM-treated peritoneal cells exhibited no specific serotonin release in response to challenge with  $10^{-6}$  M ET-1 for 15 min., versus 64% for the ET-1-challenged vehicle-treated control cells ( $P < 0.001$ ).

These results indicate that the resistance of *Kit+/+* mice (versus *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice) to ET-1-induced toxicity is largely, if not entirely, mast-cell-dependent, and that the mast cell's protective effect depends on mast-cell degranulation. Studies of rat mast cells showed that mast-cell-derived chymase and carboxypeptidase A (CPA), which are released upon mast-cell degranulation, can degrade ET-1 (ref. 14). We found that PLCs or purified PMCs from normal mice, but not PLCs from *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice, degraded ET-1 *in vitro* (Fig. 2a, b), in a time- and concentration-dependent manner (Supplementary Fig. 3), and that ET-1 degradation by *Kit+/+* PLCs was inhibited significantly by the chymase inhibitor chymostatin<sup>21</sup> but not by a CPA inhibitor<sup>14</sup>, thiorphan or leupeptin (Fig. 2c). Moreover, *Kit+/+* mice that were pretreated with chy-

mostatin developed hypothermia and diarrhoea upon injection of ET-1 i.p., and had significantly greater amounts of ET-1 in the peritoneal cavity compared with vehicle-treated, ET-1-injected *Kit+/+* mice (Fig. 2d, e). Even greater amounts of ET-1 were detected in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice injected with ET-1, but chymostatin pretreatment had no significant effects on the ET-1 levels, hypothermia, or rates of diarrhoea or death in ET-1-injected *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (Fig. 2e). These findings indicate that chymase significantly contributes to the mast cell's ability to degrade ET-1 *in vivo*, but that other mechanisms also may be involved.

We next used pharmacological and genetic approaches to characterize the mechanism of ET-1-induced mast-cell activation. In confirmation of a previous report<sup>12</sup>, we found that the stimulation of unpurified or highly purified (>95%) C57BL/6 mouse PMCs with ET-1, in concentrations as low as  $10^{-8}$  M, induced degranulation and serotonin release; moreover, ET-1 ( $10^{-7}$  M)-induced serotonin release from unpurified PMCs was greatly reduced (by 98%,  $P < 0.001$ ) by the highly selective ET<sub>A</sub> antagonist BQ-123 (ref. 22) ( $10^{-6}$  M) (Supplementary Fig. 4a, b). Pretreatment with the ET<sub>A</sub>/ET<sub>B</sub> antagonist PD142893 ( $10^{-6}$  M) (ref. 22) had an inhibitory effect that was virtually identical to that of BQ-123 (Supplementary Fig. 4c). By contrast, when compared at the same



**Figure 2** Chymase contributes to the ability of mast cells to degrade ET-1 *in vitro* and *in vivo*. PLCs or purified PMCs (>95% mast cells) from C57BL/6 mice (**a**), or PLCs derived from *Kit+/+* but not *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (**b**), can degrade ET-1; ET-1 degradation can be significantly inhibited by an inhibitor of chymase (chymostatin), but not by inhibitors of neutral endopeptidase (thiorphan), tryptase (leupeptin) or carboxypeptidase (potato CPA-inhibitor) (**c**). ET-1 ( $4 \times 10^{-6}$  M) was incubated (37 °C, 5% CO<sub>2</sub>) for 6 h (**a**) or 30 min (**b**, **c**) with PLCs ( $5.4\text{--}5.8 \times 10^6$  (**a**) or  $1.7\text{--}2.0 \times 10^6$  (**b**, **c**) total cells containing  $2 \times 10^5$  (**a**) or  $5 \times 10^4$  (**b** (in *Kit+/+* only), **c**) of PMCs), with  $2 \times 10^5$  purified PMCs (**a**) or with vehicle (**a**–**c**). Results are the percentage of control levels of ET-1 incubated with vehicle

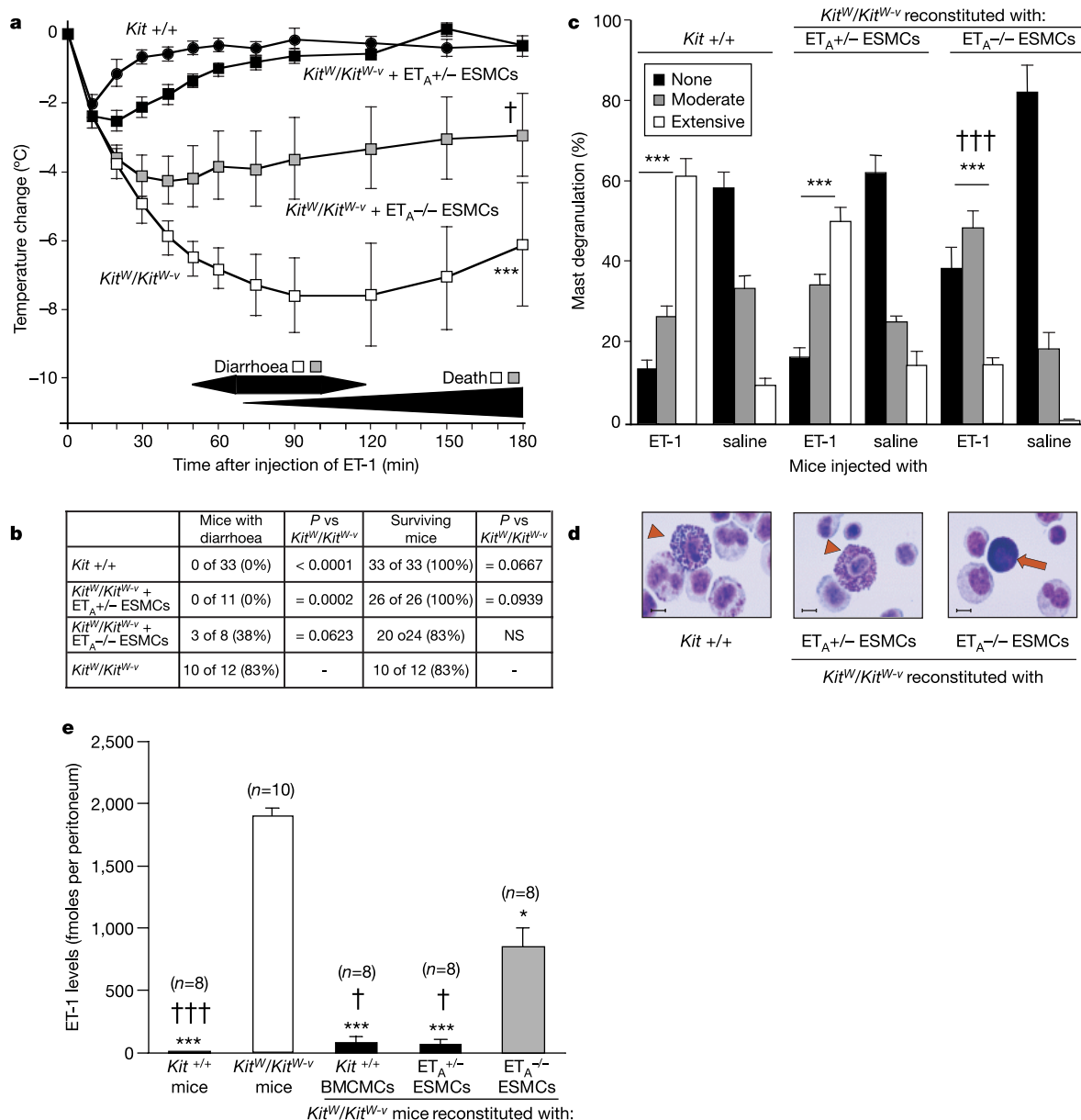
(**a**, **b**) or the percentage of ET-1 starting concentration (**c**);  $n = 3$  (**a**) or  $n = 7\text{--}9$  (**b**, **c**) per group. Mice were 6–24 weeks old. **a**–**c**, \*\*\* $P < 0.001$  versus corresponding control 'vehicle' values. **d**, **e**, 9–11-week-old female *Kit+/+* or *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice were pretreated with chymostatin (C, at 7.5 mg kg<sup>-1</sup> body weight, in vehicle, i.p.) or with vehicle alone (V = saline + 10% dimethylsulphoxide (DMSO), i.p.) and 5 min later, received an i.p. injection of ET-1 (1.2 nmol in 300  $\mu\text{l}$  saline) or saline. Body temperature, diarrhoea and survival were assessed as in Fig. 1. Amounts of intra-peritoneal ET-1 were measured at time of death or when killed (at 90 min). \*\*\* $P < 0.001$  (**d**).

concentration ( $10^{-6}$  M), the highly selective  $ET_B$  antagonist BQ-788 (ref. 22) inhibited ET-1 ( $10^{-7}$  M)-induced serotonin release to a much lesser extent (34%,  $P < 0.05$ ) (Supplementary Fig. 4d). Treatment with BQ-123 also significantly inhibited the ability of PMCs to degrade ET-1 *in vitro* (Supplementary Fig. 4e).

To assess whether the inhibition of PMC activation via  $ET_A$  could alter the susceptibility of normal mice to ET-1-induced morbidity *in vivo*, we pretreated mice with BQ-123 and then injected ET-1 i.p. (1.2 nmol) (Supplementary Fig. 5). BQ-123-pretreated  $Kit^{+/+}$  mice developed pronounced hypothermia after injection of ET-1 (albeit not to the levels observed in mast-cell-deficient  $Kit^W/Kit^{W-v}$

mice) and 13 of 15 mice developed diarrhoea, whereas the vehicle-pretreated  $Kit^{+/+}$  mice did not. Pretreatment with BQ-123 did not influence ET-1-induced hypothermia or diarrhoea in  $Kit^W/Kit^{W-v}$  mice. By contrast, responses to ET-1 after vehicle or BQ-123 pretreatment in  $Kit^{+/+}$  BMCMC-reconstituted  $Kit^W/Kit^{W-v}$  mice were virtually identical to those of  $Kit^{+/+}$  mice (compare Supplementary Figs 5c and 5a). Pretreatment with BQ-123 also markedly reduced ET-1-induced PMC degranulation in  $Kit^{+/+}$  mice, as assessed morphologically 90 min after the i.p. injection of ET-1 (Supplementary Fig. 6).

These results show that the ability of the  $ET_A$  antagonist BQ-123



**Figure 3** Genetic evidence that  $ET_A$  expression contributes to the ability of PMCs to degranulate and to reduce the levels and toxicity of ET-1 *in vivo*. **a–d**, Changes in rectal temperature (**a**), rates of diarrhoea and survival (**b**), and extent of peritoneal mast-cell degranulation (**c**, **d**) after injection of ET-1 (1.0 nmol in 200  $\mu$ l of saline, i.p.) or saline (200  $\mu$ l, i.p.) in  $Kit^{+/+}$  mice, mast-cell-deficient  $Kit^W/Kit^{W-v}$  mice and  $Kit^W/Kit^{W-v}$  mice that had been reconstituted with ESMCs that did ( $ET_A^{+/+}$ ) or did not ( $ET_A^{-/-}$ ) express  $ET_A$ . **a**, Responses of  $Kit^W/Kit^{W-v} + ET_A^{-/-}$  ESMC mice ( $n = 16$ ):  $\dagger P < 0.05$  versus  $Kit^{+/+}$  ( $n = 5$ ) or  $Kit^W/Kit^{W-v}$  ( $n = 6$ ) mice and  $P < 0.005$  versus  $Kit^W/Kit^{W-v} + ET_A^{+/+}$  ESMC mice ( $n = 15$ );  $***P < 0.005$  versus  $Kit^{+/+}$  or  $Kit^W/Kit^{W-v} + ET_A^{+/+}$  ESMC mice. **c**, Percentage of PMCs exhibiting >50%

'extensive', 10–50% 'moderate' or <10% 'none' degranulation 3 h after injection of ET-1 or saline.  $***P < 0.001$  versus corresponding saline-injected group;  $\dagger\dagger\dagger P < 0.001$  versus either of the other ET-1-injected groups. **d**, PMCs exhibiting extensive (arrowheads) or no (arrow) degranulation (Wright–Giemsa stain; scale bars, 10  $\mu$ m). Data were pooled from five (**a**, **b**) or four (**c**) experiments. **e**, Intra-peritoneal amounts of ET-1 90 min after ET-1 injection (1.2 nmol in 300  $\mu$ l saline, i.p.). Data were pooled from two experiments. ET-1 concentrations were below the detection limit (0.05 fmol ml $^{-1}$ ) in 2 of 10 mice from each group except for the  $Kit^W/Kit^{W-v}$  group (ET-1 was detectable in all these mice).  $*P < 0.05$ ,  $***P < 0.0001$  versus  $Kit^W/Kit^{W-v}$  mice;  $\dagger P < 0.05$ ,  $\dagger\dagger\dagger P < 0.001$  versus  $Kit^W/Kit^{W-v} + ET_A^{-/-}$  ESMC mice.



to enhance the hypothermia and diarrhoea induced by i.p. injection of ET-1 reflects, at least in part, its activity on mast cells. However, mast cells protected wild-type or *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice against the development of ET-1-induced mortality even in the presence of BQ-123 (boxes in Supplementary Fig. 5a, c), perhaps because a limited amount of ET<sub>B</sub>-dependent mast-cell activation can occur in this setting (Supplementary Fig. 6). To assess more directly the role of ET<sub>A</sub>-dependent mast-cell activation in protection against ET-1-induced pathology, we produced mice that contained, as their only mast-cell populations, mast cells that did or did not express ET<sub>A</sub>. Because ET<sub>A</sub><sup>-/-</sup> mice die shortly after birth<sup>23</sup> we generated mast cells *in vitro* from ET<sub>A</sub><sup>+/+</sup> or ET<sub>A</sub><sup>-/-</sup> embryonic stem (ES) cells<sup>23,24</sup>, and then adoptively transferred these ES-cell-derived mast cells (ESMCs) into *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (Fig. 3).

The numbers of PMCs in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that had been reconstituted with ET<sub>A</sub><sup>+/+</sup> versus ET<sub>A</sub><sup>-/-</sup> ESMCs were very similar ( $3.9 \pm 0.7\%$  ( $n = 15$ ) versus  $3.8 \pm 1.1\%$  ( $n = 15$ ) of PLCs, respectively), and transfer of either type of ESMCs resulted in the appearance of mast cells in the mesentery. However, the responses of these mice to an i.p. injection of ET-1 (1.0 nmol in 200  $\mu$ l of saline) were quite distinct. *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that were selectively reconstituted with ET<sub>A</sub><sup>+/+</sup> ESMCs developed slightly, but not significantly, more hypothermia than ET-1-injected *Kit<sup>+/+</sup>* mice and, like the wild-type mice, exhibited neither diarrhoea nor mortality (Fig. 3a, b); these two groups also developed very similar levels of mast-cell degranulation (Fig. 3c, d). By contrast, the *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that had been reconstituted with ET<sub>A</sub><sup>-/-</sup> ESMCs exhibited hypothermia responses and rates of diarrhoea that were intermediate between those of the *Kit<sup>+/+</sup>* mice and the mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice, and a mortality rate (17%) that was identical to that of the *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (Fig. 3a, b). In accord with these results, ET<sub>A</sub><sup>-/-</sup> ESMC-reconstituted *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice exhibited an extent of degranulation of PMCs 3 h after ET-1 injection that was significantly reduced compared with that in either group containing ET<sub>A</sub>-expressing PMCs, but that was significantly greater than that in the corresponding saline-injected mice (Fig. 3c, d).

Moreover, we found that amounts of ET-1 in the peritoneal cavities, measured 90 min after injecting ET-1 i.p. (1.2 nmol in 300  $\mu$ l of saline), were significantly higher in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice than

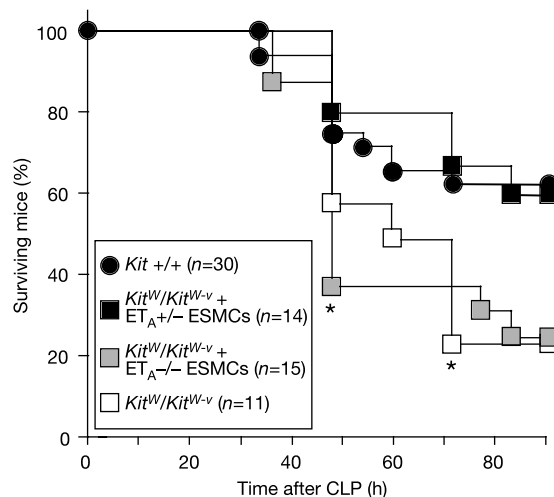
in either *Kit<sup>+/+</sup>* mice or in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice containing mast cells that expressed the ET<sub>A</sub> receptor; amounts of ET-1 intermediate between those in mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice and wild-type mice were found in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice containing mast cells that did not express ET<sub>A</sub> (Fig. 3e). These findings, and those in Fig. 2d, e, indicate that one of the mechanisms by which mast cells can limit ET-1-associated pathology *in vivo* is by directly or indirectly reducing concentrations of the peptide, and that this mast-cell function is significantly, but not entirely, ET<sub>A</sub>-dependent.

Both the data in Fig. 3 and the results in Supplementary Figs 5 and 6 indicate that the mast-cell-dependent protection against ET-1-induced toxicity observed *in vivo* may occur by both ET<sub>A</sub>-dependent and ET<sub>A</sub>-independent mechanisms. To assess the biological significance of ET<sub>A</sub>-dependent mast-cell activation during a pathological response associated with endogenous production of ET-1, we turned to the caecal ligation and puncture (CLP) model of acute bacterial peritonitis<sup>8,10,19,25</sup>. Mast cells can contribute to optimal host defence in CLP, probably through both TNF-dependent<sup>8,9,19</sup> and TNF-independent<sup>19</sup> mechanisms. Either TNF<sup>26</sup> or endotoxin<sup>27</sup> can induce ET-1 production *in vivo*, and septic peritonitis is associated with elevations in levels of ET-1 in the peritoneal cavity<sup>28</sup> and in the peripheral circulation<sup>5</sup>. Moreover, in accord with our findings in mice that had been injected i.p. with ET-1, we found that amounts of endogenous ET-1 in the peritoneal cavities, measured ~30 h after CLP, were significantly higher in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice than in either *Kit<sup>+/+</sup>* mice or in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* + *Kit<sup>+/+</sup>* BMCMC mice (Supplementary Fig. 7).

As expected based on previous work<sup>8,19</sup>, the mortality that occurred within 90 h of CLP was significantly greater in mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice than in *Kit<sup>+/+</sup>* wild-type mice (18 versus 60% survival, respectively;  $P \approx 0.043$ ; Fig. 4). Moreover, survival in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that had been reconstituted with ET<sub>A</sub><sup>+/+</sup> ESMCs (58% at 90 h) was very similar to that in wild-type mice ( $P \approx 0.96$ ) but significantly better than that in ET<sub>A</sub><sup>-/-</sup> ESMC-reconstituted *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (21%,  $P < 0.023$ ). Indeed, survival in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that had been reconstituted with ET<sub>A</sub><sup>-/-</sup> ESMCs was not statistically different ( $P \approx 0.80$ ) than that in the mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (Fig. 4).

These findings show that the expression of protective mast-cell function(s) during the CLP model of innate immunity can be promoted by ET-1- and ET<sub>A</sub>-dependent mechanisms, as well as by the interaction of bacterial products with mast-cell surface receptors (reviewed in ref. 10) or via complement activation<sup>25</sup>. Whereas other findings indicate that ET-1 can activate mast cells *in vivo* by both ET<sub>A</sub>-dependent and ET<sub>A</sub>-independent mechanisms (Fig. 3, Supplementary Fig. 6), the extent to which ET<sub>A</sub>-independent mast-cell activation reflects direct or indirect effects of ET-1 is unclear. Moreover, such mechanisms seem to be unable to compensate fully for the lack of ET<sub>A</sub> expression by mast cells during CLP (Fig. 4).

Taken together, our data indicate that ET-1 may indeed be a "two-edged sword"<sup>23</sup> in the context of CLP: helping to enhance survival via ET<sub>A</sub>-dependent mast-cell activation, which initiates a negative feedback loop that results in the degradation of ET-1 by chymase and perhaps other mast-cell-derived products, but having other effects that can contribute to pathology and mortality, particularly in the absence of ET<sub>A</sub>-dependent mast-cell activation. Whereas the specific mechanisms responsible for ET-1-induced pathology in CLP remain to be fully defined, we found that pretreatment of normal mice with the ET<sub>A</sub> antagonist, BQ-123, had no effect on survival in CLP, perhaps because BQ-123 not only partially blocked mast-cell degranulation but also inhibited adverse effects of ET-1 that are mediated by ET<sub>A</sub> receptors on other cell types (Supplementary Fig. 8a). By contrast, in ET<sub>A</sub><sup>-/-</sup> ESMC-reconstituted *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice subjected to CLP, only the effects of ET-1 on mast-cell ET<sub>A</sub> receptors were impaired (Fig. 4). However, pretreatment of wild-type mice with the ET<sub>B</sub> antagonist, BQ-788, significantly enhanced survival during CLP (Supplementary Fig. 8b). These



**Figure 4** Mast-cell expression of ET<sub>A</sub> promotes survival in CLP. Survival after CLP was enhanced in wild-type *Kit<sup>+/+</sup>* mice (filled circles) and in *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that contained adoptively transferred ET<sub>A</sub><sup>+/+</sup> ESMCs (filled squares), compared with that in mast-cell-deficient *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice (open squares) or *Kit<sup>W</sup>/Kit<sup>W-v</sup>* mice that contained ET<sub>A</sub><sup>-/-</sup> ESMCs (grey squares). \* $P < 0.05$  versus *Kit<sup>+/+</sup>* or *Kit<sup>W</sup>/Kit<sup>W-v</sup>* + ET<sub>A</sub><sup>+/+</sup> ESMC mice.

results indicate that the adverse effects of endogenous ET-1 on survival in CLP more significantly reflect ET<sub>B</sub>-dependent than ET<sub>A</sub>-dependent mechanisms. By contrast, in rats treated with ET-1 intravenously, an ET<sub>A</sub> antagonist blocked both mucosal mast-cell degranulation and ileal mucosal damage<sup>29</sup>. Thus, depending on the biological context, it is possible that ET<sub>A</sub>-dependent mast-cell activation can either promote or, via degradation of ET-1, reduce ET-1-associated pathology.

ET-1 has been implicated in a wide spectrum of pathological processes beyond sepsis, including asthma, peripheral and pulmonary hypertension, atherosclerosis, congestive heart failure, acute and chronic renal failure, gastric ulceration and cutaneous disorders, as well as in many physiological processes involving multiple organ systems<sup>2,3</sup>. Notably, mast cells also have been implicated in many of the same processes<sup>6,7,11–14,16,17,20,29</sup>. Our findings suggest that it will be of interest to investigate the potential roles of interactions between ET-1 and mast cells in these additional settings. Finally, our findings provide direct evidence for a biological function of the mast cell that has been proposed<sup>30</sup> but heretofore not proven *in vivo*: the ability to promote homeostasis by limiting the toxicity of biological substances. We speculate that ET-1 will not be the only natural compound whose toxicity can be regulated by mast-cell-dependent mechanisms. □

## Methods

### Animals

C57BL/6 mice, and genetically mast-cell-deficient WBB6F<sub>1</sub>- *Kit*<sup>W/Kit<sup>W-v</sup> (*Kit*<sup>W/Kit<sup>W-v</sup>) mice<sup>15</sup> and the congenic normal WBB6F<sub>1</sub>-+/+ (*Kit*<sup>+/+</sup>) mice, were purchased from Jackson Laboratories, Bar Harbor, Maine. Adult *Kit*<sup>W/Kit<sup>W-v</sup> mice virtually lack mature mast cells (< 0.5% the level of *Kit*<sup>+/+</sup> mice in the skin, none in the peritoneal cavity and other anatomical sites)<sup>15–17</sup>. All animal care and experimentation was conducted in accord with current National Institutes of Health and Beth Israel Deaconess Medical Center or Stanford University Institutional Animal Care and Use Committee guidelines.</sup></sup></sup>

### Caecal ligation and puncture (CLP)

CLP was performed as previously described<sup>19</sup>. Briefly, mice were deeply anaesthetized and the caecum was exposed by a 1–2 cm midline incision on the anterior abdomen and subjected to ligation of the distal half, followed by a single puncture with a 22-gauge needle. The caecum was then replaced into the abdomen, 1 ml of sterile saline was administered into the peritoneal cavity, and the incision was closed using 9-mm steel wound clips. Mice were observed for mortality at least four times daily. Mice that were clearly moribund were killed by CO<sub>2</sub> inhalation.

### Mast-cell reconstitution of *Kit*<sup>W/Kit<sup>W-v</sup> mice</sup>

Some *Kit*<sup>W/Kit<sup>W-v</sup> mice (female, 4–6-week-old) were repaired of their mast-cell deficiency selectively and locally by the injection of growth-factor-dependent BMCMCs into the peritoneal cavity<sup>18,19</sup>. Briefly, femoral bone marrow cells from *Kit*<sup>+/+</sup> mice were maintained *in vitro* for ~4 weeks in IL-3-containing medium until mast cells represented >95% of the total cells according to staining with May–Grünwald–Giemsa.  $1.0 \times 10^6$  mast cells in 200 µl of HMEM-Pipes (Hanks' minimal essential medium containing  $0.47 \text{ g l}^{-1}$  piperazine-N,N'-bis (2-ethane sulphonic acid) (Pipes buffer) instead of NaHCO<sub>3</sub>) or HMEM-Pipes alone were injected i.p. and mice were used for experiments, together with gender- and age-matched mast-cell-deficient *Kit*<sup>W/Kit<sup>W-v</sup> and *Kit*<sup>+/+</sup> mice, 4–6 weeks (Supplementary Fig. 7) or 6 months (Fig. 1a and Supplementary Fig. 5c) after adoptive transfer of cultured BMCMCs. Other *Kit*<sup>W/Kit<sup>W-v</sup> mice received, 4–6 weeks before injection with ET-1 or induction of CLP, i.p. injections of  $1.0 \times 10^6$  ESMCs that did (ET<sub>A</sub>+/-) or did not (ET<sub>A</sub>-/-) express ET<sub>A</sub>; ESMCs were generated *in vitro* from ET<sub>A</sub>+/- or -/- Strain 129 ES cells<sup>23</sup> as previously described<sup>24</sup>. The selectivity of the repair of the mast-cell deficiency of the *Kit*<sup>W/Kit<sup>W-v</sup> mice was assessed before using the mice in experiments by confirming that the adoptive transfer of BMCMCs or ESMCs failed to improve the recipients' anaemia<sup>17–19,24</sup>.</sup></sup></sup></sup>

### ET-1 measurements

ET-1 levels were measured by ELISA (Biomedica, Vienna).

### Statistical analysis

The differences in percentages of mice exhibiting diarrhoea or death after injection with ET-1 were compared by Fisher's exact test, whereas analysis of variance (ANOVA) for repeated measures was used to assess differences in body temperature responses to ET-1. The significance of differences in the survival rates after CLP was assessed using the Mantel–Cox Logrank test and the significance of differences in percentages of peritoneal mast-cell degranulation was assessed by the  $\chi^2$  test. All other data were tested for statistical significance using the unpaired two-tailed Student's *t*-test or Mann–Whitney *U*-test; NS, not significant (*P* > 0.05). Unless otherwise specified, all data are presented as mean ± s.e.m.

Received 30 July; accepted 6 October 2004; doi:10.1038/nature03085.

Published online 14 November 2004.

- Yanagisawa, M. *et al.* A novel potent vasoconstrictor peptide produced by vascular endothelial cells. *Nature* **332**, 411–415 (1988).
- Wanecek, M., Weitzberg, E., Rudehill, A. & Oldner, A. The endothelin system in septic and endotoxin shock. *Eur. J. Pharmacol.* **407**, 1–15 (2000).
- Kedzierski, R. M. & Yanagisawa, M. Endothelin system: the double-edged sword in health and disease. *Annu. Rev. Pharmacol. Toxicol.* **41**, 851–876 (2001).
- Szalai, L., Kaszaki, J., Nagy, S. & Boros, M. The role of endothelin-1 in circulatory changes during hypodynamic sepsis in the rat. *Shock* **10**, 123–128 (1998).
- Ornan, D. A., Chaudry, I. H. & Wang, P. The dissociation between upregulated endothelins and hemodynamic responses during polymicrobial sepsis. *Biochim. Biophys. Acta* **1501**, 211–218 (2000).
- Kawakami, T. & Galli, S. J. Regulation of mast-cell and basophil function and survival by IgE. *Nature Rev. Immunol.* **2**, 773–786 (2002).
- Benoist, C. & Mathis, D. Mast cells in autoimmune disease. *Nature* **420**, 875–878 (2002).
- Echtenacher, B., Männel, D. N. & Hültner, L. Critical protective role of mast cells in a model of acute septic peritonitis. *Nature* **381**, 75–77 (1996).
- Malaviya, R., Ikeda, T., Ross, E. & Abraham, S. N. Mast cell modulation of neutrophil influx and bacterial clearance at sites of infection through TNF- $\alpha$ . *Nature* **381**, 77–80 (1996).
- Galli, S. J., Chatterjea, D. & Tsai, M. in *The Innate Immune Response to Infection* (ed. Gordon, S.) 111–132 (ASM, Berlin, 2004).
- Ehrenreich, H. *et al.* Endothelins belong to the assortment of mast cell-derived and mast cell-bound cytokines. *New Biol.* **4**, 147–156 (1992).
- Yamamura, H., Nabe, T., Kohno, S. & Ohata, K. Endothelin-1, one of the most potent histamine releasers in mouse peritoneal mast cells. *Eur. J. Pharmacol.* **265**, 9–15 (1994).
- Yamamura, H., Nabe, T., Kohno, S. & Ohata, K. Mechanism of histamine release by endothelin-1 distinct from that by antigen in mouse bone marrow-derived mast cells. *Eur. J. Pharmacol.* **288**, 269–275 (1995).
- Metsärinne, K. P. *et al.* Activated mast cells increase the level of endothelin-1 mRNA in cocultured endothelial cells and degrade the secreted peptide. *Arterioscler. Thromb. Vasc. Biol.* **22**, 268–273 (2002).
- Kitamura, Y., Go, S. & Hatanaka, K. Decrease of mast cells in W/W<sup>v</sup> mice and their increase by bone marrow transplantation. *Blood* **52**, 447–452 (1978).
- Galli, S. J., Zsebo, K. M. & Geissler, E. N. The kit ligand, stem cell factor. *Adv. Immunol.* **55**, 1–96 (1994).
- Tsai, M., Tam, S. Y., Wedemeyer, J. & Galli, S. J. Mast cells derived from embryonic stem cells: a model system for studying the effects of genetic manipulations on mast cell development, phenotype, and function *in vitro* and *in vivo*. *Int. J. Hematol.* **75**, 345–349 (2002).
- Nakano, T. *et al.* Fate of bone marrow-derived cultured mast cells after intracutaneous, intraperitoneal, and intravenous transfer to genetically mast cell-deficient W/W<sup>v</sup> mice. Evidence that cultured mast cells can give rise to both connective tissue and mucosal mast cells. *J. Exp. Med.* **162**, 1025–1043 (1985).
- Maurer, M. *et al.* The c-kit ligand, stem cell factor, can enhance innate immunity through effects on mast cells. *J. Exp. Med.* **188**, 2343–2348 (1998).
- Boesiger, J. *et al.* Mast cells can secrete vascular permeability factor/vascular endothelial cell growth factor and exhibit enhanced release after immunoglobulin E-dependent upregulation of Fc $\epsilon$  receptor 1 expression. *J. Exp. Med.* **188**, 1135–1145 (1998).
- Maguire, J. J., Kuc, R. E. & Davenport, A. P. Vasoconstrictor activity of novel endothelin peptide, ET-1<sub>(1–31)</sub>, in human mammary and coronary arteries *in vitro*. *Br. J. Pharmacol.* **134**, 1360–1366 (2001).
- Opgenorth, T. J. Endothelin receptor antagonism. *Adv. Pharmacol.* **33**, 1–64 (1995).
- Clouthier, D. E., Williams, S. C., Hammer, R. E., Richardson, J. A. & Yanagisawa, M. Cell-autonomous and nonautonomous actions of endothelin-A receptor signaling in craniofacial and cardiovascular development. *Dev. Biol.* **261**, 506–519 (2003).
- Tsai, M. *et al.* *In vivo* immunological function of mast cells derived from embryonic stem cells: an approach for the rapid analysis of even embryonic lethal mutations in adult mice *in vivo*. *Proc. Natl Acad. Sci. USA* **97**, 9186–9190 (2000).
- Prodeus, A. P., Zhou, X., Maurer, M., Galli, S. J. & Carroll, M. C. Impaired mast cell-dependent natural immunity in complement C3-deficient mice. *Nature* **390**, 172–175 (1997).
- Ahn, G. Y. *et al.* The expression of endothelin-1 and its binding sites in mouse skin increased after ultraviolet B irradiation or local injection of tumor necrosis factor  $\alpha$ . *J. Dermatol.* **25**, 78–84 (1998).
- Sugiura, M., Inagami, T. & Kon, V. Endotoxin stimulates endothelin-release *in vivo* and *in vitro* as determined by radioimmunoassay. *Biochem. Biophys. Res. Commun.* **161**, 1220–1227 (1989).
- Lundblad, R. & Giercksky, K. E. Endothelin concentrations in experimental sepsis: profiles of big endothelin and endothelin 1–21 in lethal peritonitis in rats. *Eur. J. Surg.* **161**, 9–16 (1995).
- Boros, M., Szalai, L. & Kaszaki, J. Endothelin-1 induces mucosal mast cell degranulation and tissue injury via ET<sub>A</sub> receptors. *Clin. Sci. (Lond.)* **103** (suppl. 48), 31S–34S (2002).
- Higginbotham, R. D. Mast cells and local resistance to Russell's viper venom. *J. Immunol.* **95**, 867–875 (1965).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank R. Parker of the Biometrics Center of the Beth Israel Deaconess Medical Center for consultation regarding the statistical analysis of the data, S. Fish, L. Fox, M. Liebersbach and A. Xu for technical assistance, M.-H. Jouvin for discussions and critical reading of the manuscript, and R. Paus for supporting M. Metz. This work was supported by United States Public Health Science Grants (to S.J.G.), by grants of the Deutsche Forschungsgemeinschaft (to M. Maurer, J.W. and M. Metz), and by a grant of the Boehringer Ingelheim Fonds (to M. Metz).

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to S.J.G. ([sgalli@stanford.edu](mailto:sgalli@stanford.edu)).

# The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II

Minkyu Kim<sup>1</sup>, Nevan J. Krogan<sup>2</sup>, Lidia Vasiljeva<sup>1</sup>, Oliver J. Rando<sup>3</sup>, Eduard Nedea<sup>2</sup>, Jack F. Greenblatt<sup>2</sup> & Stephen Buratowski<sup>1</sup>

<sup>1</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

<sup>2</sup>Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, Ontario, Canada M5G 1L6

<sup>3</sup>Bauer Center for Genomics Research, Harvard University, 7 Divinity Ave., Cambridge, Massachusetts 02138, USA

The carboxy-terminal domain (CTD) of the RNA polymerase II (RNAPII) largest subunit consists of multiple heptapeptide repeats with the consensus sequence YSPTSPS. Different CTD phosphorylation patterns act as recognition sites for the binding of various messenger RNA processing factors, thereby coupling transcription and mRNA processing<sup>1</sup>. Polyadenylation factors are co-transcriptionally recruited by phosphorylation of CTD serine 2 (ref. 2) and these factors are also required for transcription termination<sup>3,4</sup>. RNAPII transcribes past the poly(A) site, the RNA is cleaved by the polyadenylation machinery, and the RNA downstream of the cleavage site is degraded. Here we show that Rtt103 and the Rat1/Rai1 5' → 3' exonuclease are localized at 3' ends of protein coding genes. In *rat1-1* or *rai1Δ* cells, RNA 3' to polyadenylation sites is greatly stabilized and termination defects are seen at many genes. These findings support a model in which poly(A) site cleavage and subsequent degradation of the 3'-downstream RNA by Rat1 trigger transcription termination<sup>5,6</sup>.

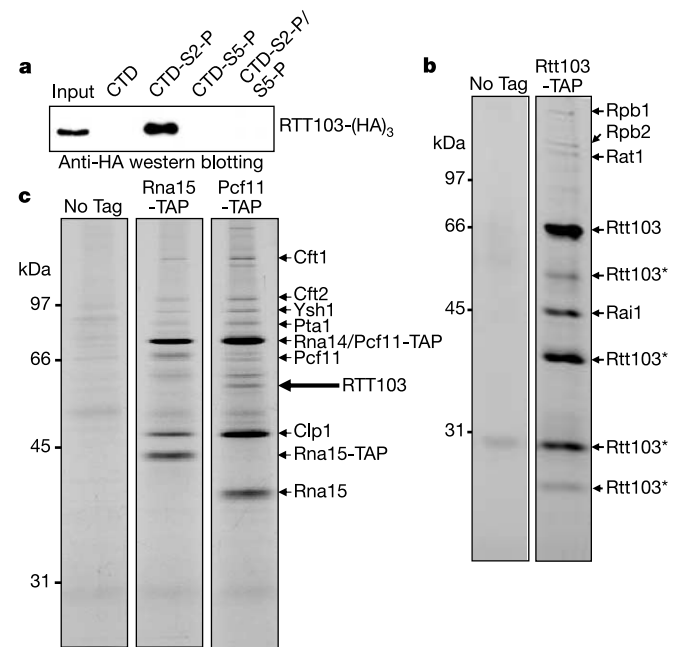
CTD peptides, consisting of four heptapeptide repeats with different phosphorylation states, were immobilized on beads and used to isolate novel CTD-binding proteins by affinity chromatography of yeast whole-cell extracts. A 60-kDa band that specifically bound to CTD phosphorylated at serine 2 (ser2-P) was identified as Rtt103/YDR289c by mass spectrometry (data not shown). Specific binding of haemagglutinin (HA)-tagged Rtt103 to the ser2-P CTD peptide was verified by immunoblotting (Fig. 1a). Rtt103 was originally isolated as a regulator of Ty1 transposition, because Ty1 RNA synthesis and transposition increase slightly in a *rtt103Δ* strain<sup>7</sup>. Sequence analysis of Rtt103 predicts that it has an RPR domain (also known as a CTD-interacting domain or CID) present in several proteins involved in regulation of nuclear pre-mRNA<sup>8</sup>. *Saccharomyces cerevisiae* has two other RPR/CID proteins, Pcf11 and Nrd1, both of which are involved in mRNA and small nucleolar RNA (snoRNA) 3'-end processing and both of which interact with phosphorylated CTD (ref. 9 and references therein). Rtt103 has homologues in higher eukaryotes, and a mouse homologue (BC021395/Q8VD54) was found in a phospho-CTD immunoprecipitation (S. McCracken and B. Blencowe, personal communication).

An Rtt103 deletion strain is viable. Rtt103 function was probed by synthetic genetic array (SGA) analysis<sup>10</sup>. The *rtt103Δ* strain was crossed to a set of viable deletion strains selected for their involvement in gene expression, and the resulting double-mutant spores were analysed (data not shown). Synthetic phenotypes were confirmed by tetrad dissections. Genetic interactions suggest a role for Rtt103 in 3'-end processing. Synthetic lethality of *rtt103Δ* was observed when combined with a deletion of *CTK1* and synthetic slow growth with a *CTK2* deletion. Ctk1 and Ctk2 are components of CTDK-1 (ref. 11), the kinase responsible for CTD ser2 phosphorylation and co-transcriptional recruitment of polyadenylation factors<sup>2,12</sup>. In addition, a double knockout of *REF2* and *RTT103* is

non-viable. Ref2 is involved in snoRNA and mRNA 3'-end formation<sup>13,14</sup>. Synthetic slow growth phenotypes were seen with mutants of several transcription factors implicated in elongation or 3' processing. These include Spt4, the Elongator complex (ELP1-6), members of the Paf complex, Bur2, Htz1 and its assembly complex SWR-C, and several ubiquitin/proteasome-related proteins. Finally, synthetic lethality was observed with a deletion of *RAI1*, which encodes a component of an RNA degradation complex (see below).

Tandem affinity purification (TAP) of Rtt103 was used to identify associated proteins. In addition to the expected subunits of RNAPII, the Rat1 and Rai1 proteins co-purified with Rtt103 (Fig. 1b). Rat1 (also named Xrn2 and Hke1) is a nuclear 5' → 3' exoRNase involved in 5.8S RNA and snoRNA 5' trimming<sup>15,16</sup>. Rai1 is a non-essential protein that co-purifies with Rat1 and enhances Rat1 activity *in vitro*<sup>17</sup>. Rat1/Rai1 has not previously been directly implicated in mRNA processing, but nuclear export of poly(A)<sup>+</sup> RNA is blocked in *rat1-1* temperature-sensitive cells at the non-permissive temperature<sup>18</sup>. This nuclear retention phenotype is also seen in cells mutated in some polyadenylation factors<sup>19</sup>. Interestingly, Rtt103 was also isolated in TAP purifications of cleavage factor Pcf11 when washing was performed at lower stringency, but not in purifications of Rna15 or other polyadenylation factors (Fig. 1c, ref. 14, and data not shown).

To determine whether Rtt103, Rat1 and Rai1 associate with transcription elongation complexes, chromatin immunoprecipitation (ChIP) was used to monitor their positions on the *ADH1*, *PMA1* and *PYK1* genes (Fig. 2a–c). Rtt103 crosslinked very strongly near 3' ends of genes, a pattern similar to that of polyadenylation factors such as Rna14, Rna15 and Pcf11 (refs 2, 4). Consistent with its ser2-P CTD binding and interactions with polyadenylation



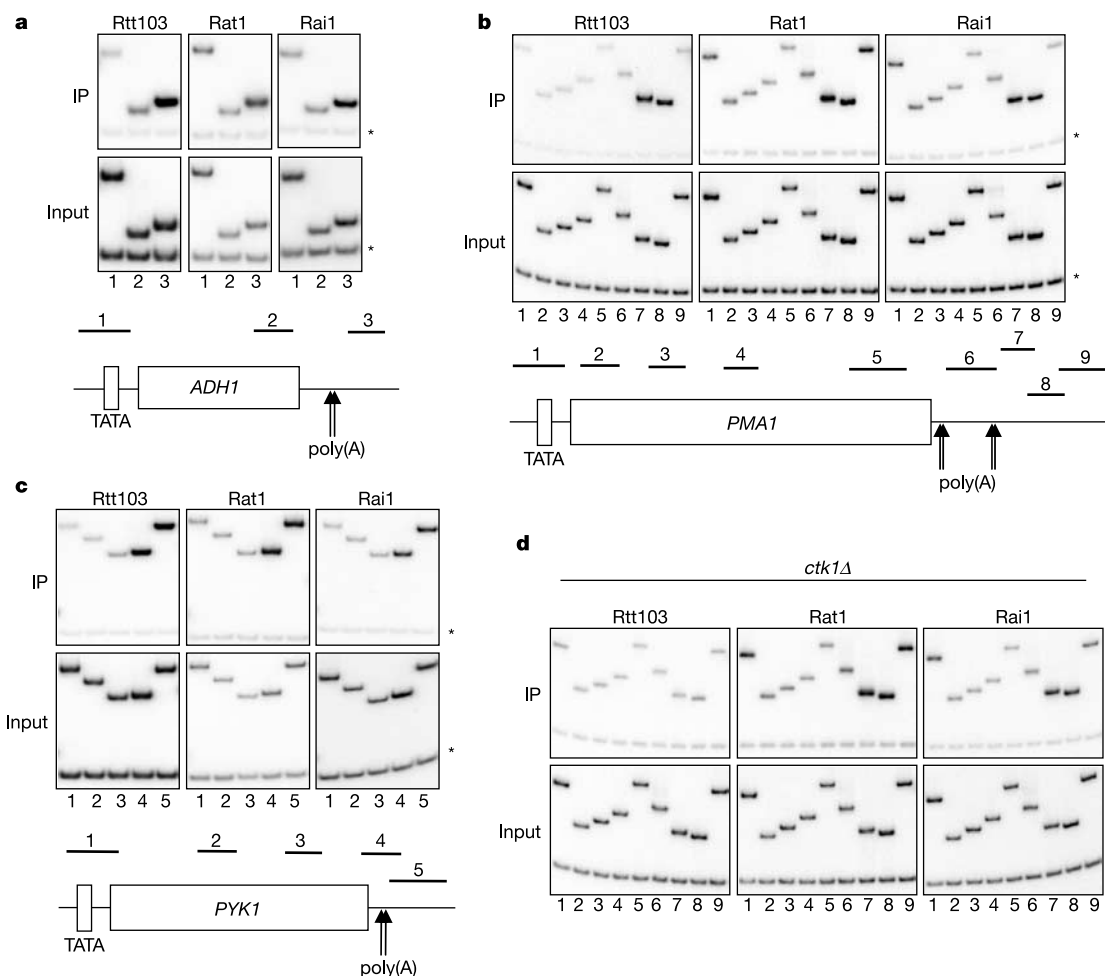
**Figure 1** Isolation of a complex containing Rtt103, RNA polymerase II and the Rat1/Rai1 exonuclease. **a**, Rtt103 specifically interacts with CTD phosphorylated at serine 2. Yeast extracts from a triple haemagglutinin (HA)<sub>3</sub>-tagged Rtt103 strain were incubated with beads carrying CTD peptides with the indicated phosphorylations. After washing, bound Rtt103 was monitored by immunoblot analysis with anti-HA antibody (12CA5). **b**, Purification of proteins associated with TAP-tagged Rtt103. Proteins were analysed by SDS-PAGE and silver staining, and identified by mass spectrometry. Asterisks indicate degradation products of Rtt103. **c**, Purification and mass spectroscopy analysis of TAP-tagged Rna15 and Pcf11 were performed as in **b**.



factors was our failure to observe recruitment of Rtt103 to *PMA1* or other genes in a *ctk1Δ* strain (Fig. 2d; data not shown). Crosslinking of Rat1 and Rai1 was also strongest at 3' ends of genes, indicating a possible role in 3'-end processing. Rat1 and Rai1 also showed weaker crosslinking to promoter and coding regions. Interestingly, Rat1/Rai1 crosslinking was not dependent on Ctk1. Furthermore, Rat1 crosslinking was the same in *rtt103Δ* or *rai1Δ* strains (data not shown). Therefore, either Rat1 can intrinsically recognize cleaved mRNA 5' ends or there are redundant mechanisms for targeting the nuclease. Redundancy could explain the synthetic lethality observed between *RAI1* and *RTT103* deletions.

What is the role of Rat1/Rai1 on mRNA genes? At 5' ends and coding regions, Rat1 might act to degrade any uncapped mRNAs. This would be one of many 'quality control' surveillance mechanisms that prevent incomplete mRNAs from being transported and translated<sup>20</sup>. In support of this idea, Rat1/Rai1 contributes to increased nuclear mRNA degradation<sup>21,22</sup>. At 3' ends of genes, RNAPII transcribes past polyadenylation sites. Cleavage generates a new, uncapped 5'-RNA end on the downstream RNA. A coupled *in vitro* termination/polyadenylation system showed rapid degradation of this downstream RNA, which does not become part of the mature transcript<sup>23</sup>. Because Rat1 is a nuclear 5' → 3' exonuclease, we suspected it might be responsible for the degradation of the downstream RNA.

To test this possibility, total RNA was reverse transcribed with a primer complementary to the 3'-downstream region of the *ADH1* gene (Fig. 3a). This primer can produce two complementary DNAs: one from uncleaved pre-mRNA transcripts and one from the downstream RNA after cleavage. Polymerase chain reaction (PCR) was then performed with three primer pairs to generate products 1, 2 and 3, respectively (Fig. 3a). In wild-type cells, the amounts of the three PCR products were similar, indicating that most of the signal was from uncleaved pre-mRNA (Fig. 3b, left panel). In contrast, *rat1-1* cells at both the permissive and non-permissive temperatures had 10–20-fold higher levels of PCR product 3 (Fig. 3b, middle panel). Note that *rat1-1* is also partly defective in its snoRNA-processing functions at permissive temperatures<sup>15</sup>. Little increase was seen in products 1 and 2, suggesting that the increased product 3 is due to stabilization of the downstream RNA and not defects in mRNA 3' cleavage. Rai1 enhances Rat1 activity *in vitro*<sup>17</sup> and a *rai1Δ* strain also showed stabilization of downstream RNA (Fig. 3b, right panel). A smaller but reproducible increase of product 3 was also seen with *rtt103Δ*, although some increase in product 2 was also observed. Similar results were obtained for the *PMA1* gene (data not shown). These results indicate that Rat1, aided by Rai1 and perhaps Rtt103, is the enzyme responsible for degrading the 3'-downstream RNA generated after poly(A) site cleavage.



**Figure 2** Rtt103 and Rat1/Rai1 localize at 3' ends of genes. ChIP was performed with strains containing TAP-tagged Rtt103, Rat1 or Rai1. Sheared chromatin was precipitated with IgG-agarose and then amplified by PCR (upper panel) with primers as shown in the diagrams at the bottom. The upper band is the gene-specific band; the common lower band (asterisk) is an internal background control from a non-transcribed region on

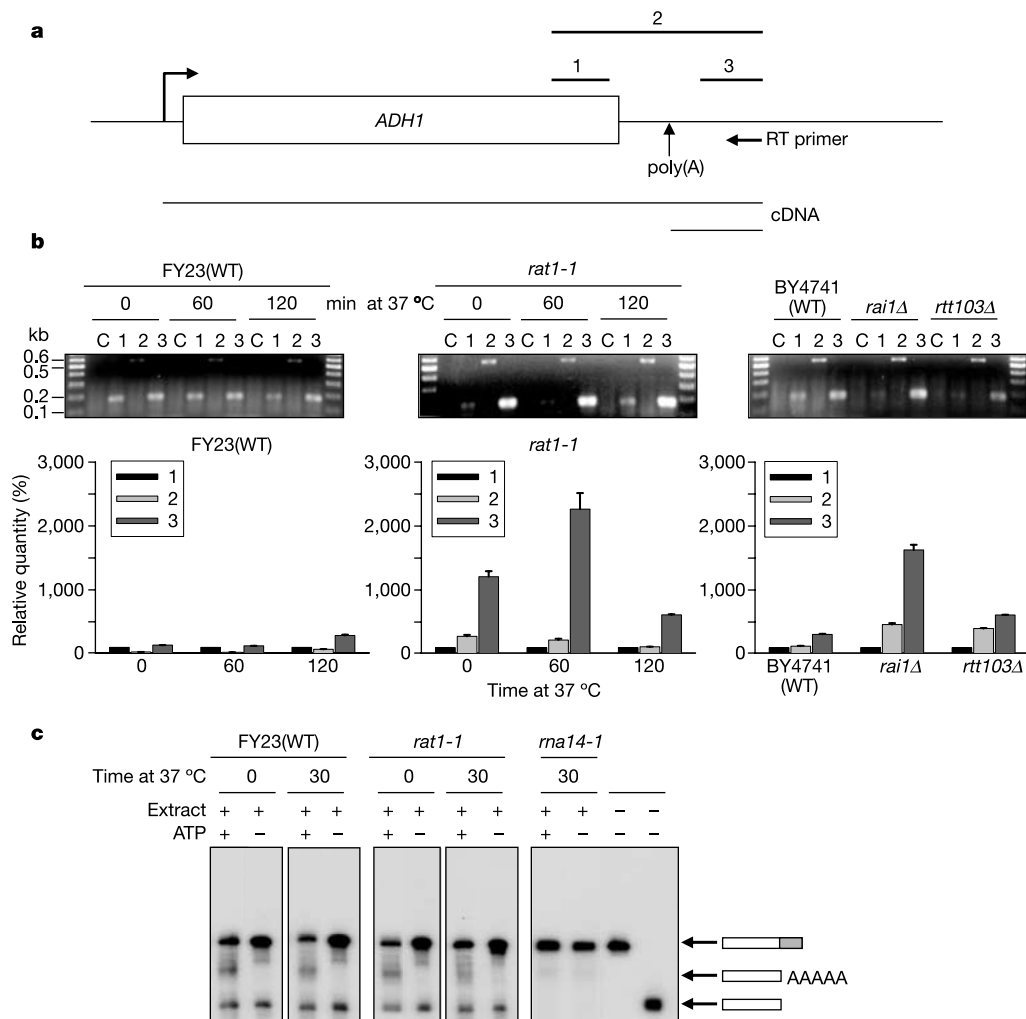
chromosome V. The middle panels show input controls used to normalize the PCR amplification efficiency of each primer pair. **a–c**, Analysis was performed on *ADH1* (**a**) *PMA1* (**b**) and *PYK1* (**c**). **d**, ChIP analysis on the *PMA1* gene in a *ctk1Δ* strain. IP, immunoprecipitation.

The *RAT1* gene was initially isolated in a screen for mutants defective in poly(A)<sup>+</sup> RNA transport out of the nucleus<sup>18</sup>. Many polyadenylation factor mutants exhibit a similar nuclear retention phenotype<sup>19</sup>. Extracts from a *rat1-1* mutant strain were assayed for transcript cleavage and polyadenylation *in vitro*, but no defects were observed (Fig. 3c and Supplementary Fig. 1a). To determine whether Rat1/Rai1 instead has a function in transcription termination, ChIP analysis of RNAPII was performed in *rai1Δ* and *rat1-1* strains. In wild-type cells, Rpb3 (that is, RNAPII) density is roughly even throughout the transcribed region of the *PMA1* gene, but decreases to near background about 200–400 base pairs downstream of the polyadenylation site (Fig. 4a, see primers 8 and 9). However, in *rat1-1* cells at both 23 °C and 37 °C, RNAPII density did not decrease, indicating a possible role for Rat1 in termination. Cells lacking Rai1 were similar to *rat1-1* cells at 23 °C, but *rtt103Δ* cells had no obvious termination defect (Fig. 4b).

To see how general the 3' effect was, immunoprecipitated chromatin was hybridized to a tiled oligonucleotide microarray covering *S. cerevisiae* chromosome III. Input and anti-Rpb3-immunoprecipitated DNAs were amplified, labelled with Cy3 or Cy5 and hybridized to the arrays. In *rat1-1* cells, most genes in which

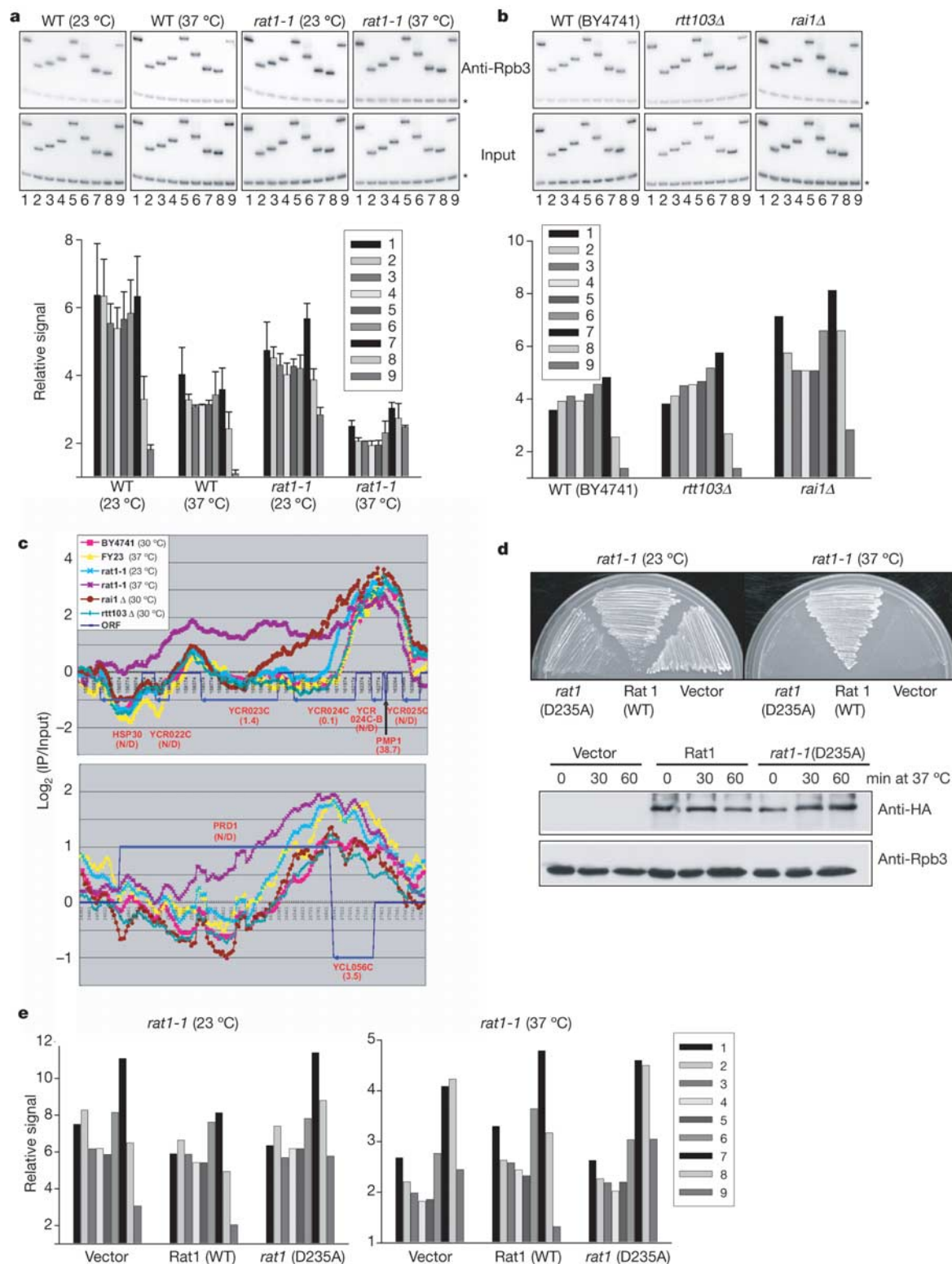
termination could be seen exhibited a 3' extension of Rpb3 crosslinking relative to that observed in wild-type cells. Two such genes are shown in Fig. 4c. The effect was most obvious on strongly transcribed genes well separated from other strong transcription units. For example, the *PMP1* gene (transcription rate 38.7 mRNA h<sup>-1</sup>; see ref. 24), which is upstream of *YCR024C* (0.1 mRNA h<sup>-1</sup>), shows a steep peak of Rpb3 crosslinking in wild-type cells. In the *rai1Δ* strain, RNAPII was seen much further downstream, declining at a more gradual rate. The *rat1-1* strain also showed extended 3'-RNAPII crosslinking at 23 °C. At 37 °C, transcription in this strain was markedly extended several kilobases farther downstream (Fig. 4c). Thus, Rat1 probably has a function in transcription termination at multiple loci throughout the genome.

Two experiments were performed to distinguish whether the Rat1 exonuclease itself, or just the presence of the protein, was required for termination. First, immunoblotting of the Rat1-1 protein showed that it was stable at all temperatures (Supplementary Fig. 1b), showing that the termination defect was not simply due to an absence of Rat1. Next, a point mutant of Rat1 was created (D235A) in a critical residue that is absolutely conserved in this nuclease family<sup>25</sup>. The mutant produced normal levels of Rat1



**Figure 3** Rat1 degrades 3'-downstream RNA but is not required for mRNA cleavage or polyadenylation. **a**, Schematic diagram of *ADH1*. Lines below the gene indicate potential cDNAs. Bars above denote RT-PCR products shown in **b**. **b**, RT-PCR with RNA from wild-type (WT, left panel) and *rat1-1* cells (middle) shifted to 37 °C for the indicated duration. A similar analysis of *rai1Δ*, *rtt103Δ* and WT cells is shown at the right. C is a no-RT control. For quantification (below each gel), product 1 was set to 100% in each

group. Results are means  $\pm$  s.d. for three repetitions. Black bars, product 1; light grey bars, product 2; dark grey bars, product 3. **c**, *In vitro* cleavage/polyadenylation assays were performed with extracts from the indicated strains. The last two lanes show standards for precursor RNA and cleavage product. The *rna14-1* extract was competent for polyadenylation of pre-cleaved product (Supplementary Fig. 1a).



**Figure 4** Rat1/Rai1 promotes transcription termination. **a**, ChIP analysis of Rpb3 on *PMA1* in wild-type (WT) and *rat1-1* strains grown at the indicated temperature. PCR products are shown as in Fig. 2b. Quantification is below each gel. The y-axis shows the ratio of *PMA1* signal relative to the internal negative control. A value of 1.0 therefore indicates no signal above background. Error bars show standard deviation from three repetitions. **b**, The same analysis with isogenic WT, *rai1Δ* and *rtt103Δ* strains. **c**, Chromatin immunoprecipitated with anti-Rpb3 was analysed using microarrays as described in Methods. Two representative genes are displayed here. The y-axis is log<sub>2</sub> of the immunoprecipitation (IP) to Input ratio, and the x-axis represents position along the

chromosome. Dark blue arrows denote open reading frames (ORFs). Note that, for the genes of interest, transcription is right to left. The numbers in parentheses show reported transcription rates (mRNA h<sup>-1</sup>), with N/D being below the detection limit<sup>24</sup>. **d**, Rat1 (D235A) does not support viability. Plasmids containing no insert, WT *RAT1* or the mutant D235A were transformed into a *rat1-1* strain and tested for growth at 23 and 37 °C. Below the plates are immunoblots showing expression of the tagged wild-type and D235A Rat1 proteins. **e**, Wild-type *RAT1* gene restores termination, whereas the D235A mutant does not. Cells from **d** were used for ChIP experiments as in **a**.



protein but was inactive *in vitro* (Supplementary Fig. 2) and failed to complement the temperature sensitivity of the *rat1-1* strain (Fig. 4d). Whereas wild-type *RAT1* corrected the termination defect of the *rat1-1* strain, the D235A mutant did not (Fig. 4e). In fact, the point mutant caused slightly slower growth and exacerbated the read-through phenotype at the permissive temperature. The exonuclease activity of Rat1 is therefore required for proper termination.

How does Rat1 contribute to transcription termination? The most obvious mechanism is by degrading the transcript behind elongating RNAPII. On reaching the polymerase, Rat1 might trigger transcript release by disrupting polymerase–RNA contacts. This mechanism echoes rho-dependent termination in bacteria. Although rho does not degrade RNA, it tracks along the RNA behind polymerase and triggers termination on reaching the elongation complex. Alternatively, degrading upstream RNA could promote ‘backtracking’ and arrest by RNAPII (ref. 26). Without RNA behind the elongation complex, backtracking polymerases might release from the DNA template.

A cleavage-dependent tracking/termination model has been called the ‘torpedo’ model<sup>5,6</sup>, and our results provide support for this idea. An alternative ‘anti-terminator’ model proposes that emergence of the polyadenylation site changes the properties of the elongation complex independently of cleavage to trigger termination, perhaps by dissociation of a positive elongation factor or recruitment of a termination factor. It is likely that termination involves aspects of both models. Association of the Paf and TREX/THO elongation factors with transcribing RNAPII is reduced on passage through a poly(A) site<sup>2,4</sup>, perhaps promoting termination. Polyadenylation factors are recruited to 3′ ends of genes dependent on CTD ser2 phosphorylation and the poly(A) site<sup>2,4</sup>, leading to mRNA cleavage and degradation of the downstream RNA by Rat1. The role of Rat1 in termination seems to be conserved over evolution, because an RNAi knockdown of the mammalian Rat1 homologue (*XRN2*) also results in termination defects<sup>27</sup>. □

## Methods

### Yeast strains

Strains used are listed in Supplementary Table 1.

### CTD peptide binding assay

CTD peptides were synthesized with four repeats of YSPTSPS and amino-terminal biotinylation. Unphosphorylated, ser2, ser5 and ser2/ser5 phosphorylated peptides were made. For the identification of ser2-P-specific binding proteins, 3 mg of streptavidin-coated magnetic beads (M-280; Dynal) was resuspended in 450 µl of PBS, mixed with 30 µg of ser2-P peptide and incubated at 4 °C for 90 min. After being washed twice with PBS, the beads were resuspended in 450 µl of PBS, mixed with 3 µg of unphosphorylated CTD peptide and incubated again for 90 min. Beads with the other peptides were prepared similarly. Beads were washed twice with PBS and resuspended in 900 µl of binding buffer (10 mM potassium phosphate pH 7.7, 100 mM potassium acetate, 20 mM magnesium acetate, 5 mM EGTA, 10% glycerol, 0.1% Nonidet P40, 0.05% Triton X-100, 1 mM dithiothreitol, 1 mM Na<sub>2</sub>S<sub>2</sub>O<sub>3</sub>, 1 mM NaF, 0.4 mM NaVO<sub>3</sub>) plus protease inhibitors. Yeast whole cell extract (5 mg) was added and the mixture was incubated at 4 °C overnight. After washing of the beads four times with 450 µl of binding buffer, bound proteins were eluted with 40 µl of binding buffer in the presence of 1 M potassium acetate. The eluate was dialysed twice for 1 h against 200 ml of binding buffer at 4 °C, with Millipore type VS 0.025 µm filter and subjected to SDS–polyacrylamide-gel electrophoresis (SDS–PAGE). Protein bands were compared between unphosphorylated and ser2-P-affinity purification, and specific bands were analysed by mass spectrometry. For immunoblotting, protein-bound beads were boiled and loaded on SDS–PAGE gels directly. For Fig. 1a, a similar reaction was performed at a smaller scale with yeast extract from strain YSB815, in which Rtt103 is tagged with HA. Bound proteins were analysed by SDS–PAGE and anti-HA (monoclonal 12CA5) immunoblotting.

### Rat1 mutants

The *rat1-1* mutant strain was from C. Cole (Dartmouth University, New Hampshire), and Rat1 plasmids were from A. Johnson (University of Texas, Austin). Rat1 and Rat1-1 were TAP-tagged as described previously<sup>14,28</sup>. The D235A mutant was constructed by PCR mutagenesis and cloned into a vector in which the HA-tagged protein was expressed from the *RAT1* promoter. A similar wild-type construct was made in parallel.

### TAP purification

TAP purifications at 100 mM NaCl were performed as described<sup>14,28</sup>.

### SGA analysis

SGA analysis was performed as described previously<sup>10</sup>.

### ChIP

ChIP procedures and quantification were performed as described<sup>4</sup>. For the experiment in Fig. 4, both wild-type and *rat1-1* cells were incubated at 23–25 °C until OD<sub>595</sub> reached 0.8. Equal amounts of medium prewarmed to 51 °C were added and cells were incubated at 37 °C for a further 25–30 min. Formaldehyde crosslinking and chromatin preparation were as described.

### Polyadenylation and cleavage assays

RNA substrates containing the Gal7 3′ end were synthesized *in vitro* and assayed as described previously<sup>29</sup>.

### Reverse transcription (RT)–PCR analysis

Total RNA was isolated by extraction with hot phenol from isogenic FY23(WT) and *rat1-1* cells at 0, 60 and 120 min after temperature shift to 37 °C, or from isogenic BY4741(WT), *Δrtt103* and *Δrail* cells at 30 °C. About 3 µg of total RNA was reverse transcribed. Priming was with 10 pmol of an anti-sense oligonucleotide spanning the 3′-downstream region of the *ADH1* gene; 5′-CCCACTGAAGGCTAGGCTGTGG-3′. One-tenth of the reverse-transcribed cDNA products were then amplified by PCR. RT–PCR products were subjected to electrophoresis on 1.2% agarose gel and revealed by ethidium bromide staining. For each set of RT–PCR reactions, the relative intensities of PCR products 2 and 3 were calculated by setting that of PCR product 1 arbitrarily to 100%.

### Microarray experiments

Formaldehyde crosslinking and ChIP were performed as above. Immunoprecipitated DNAs were amplified as described in [www.microarrays.org/pdfs/Round\\_A\\_B\\_C.pdf](http://www.microarrays.org/pdfs/Round_A_B_C.pdf) with some modifications. Control and immunoprecipitated DNAs from 30 ml of cells were randomly tagged with Sequenase and primer A (5′-GTTTCCCACTGACGATCNNNNNNNN-3′) and then further amplified by PCR with primer B (5′-GTTTCCCACTGACGATC-3′). Finally, 2 µg of amplified DNA was labelled by Klenow fragment in the presence of Cy dye-coupled dCTP and random primers (BioPrime labelling kit; Invitrogen). Input and immunoprecipitated DNA were labelled with different dyes for competitive hybridization to spotted microarrays<sup>30</sup> consisting of 50-base oligonucleotides spanning *S. cerevisiae* chromosome III (O.J.R., unpublished), tiled every 20 base pairs. Hybridization was at 55 °C for 4 h in buffer containing 3.4 × SSC and 0.3% SDS. Arrays were washed sequentially with 1 × SSC/0.03% SDS, 0.2 × SSC and 0.05 × SSC. Microarrays were scanned and fluorescence intensities were quantified with an Axon GenePix 4000B scanner and software. Data in Fig. 4c are from a moving-window average of five probes. Hybridizations with BY4741 and the *rat1-1* were repeated several times, whereas the FY23, *railΔ* and *rtt103Δ* strains were tested once. A full analysis of this and other data will be presented elsewhere (M.K., S.B. and O.J.R., unpublished work).

Received 15 June; accepted 17 September 2004; doi:10.1038/nature03041.

- Buratowski, S. The CTD code. *Nature Struct. Biol.* **10**, 679–680 (2003).
- Ahn, S. H., Kim, M. & Buratowski, S. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3′ end processing. *Mol. Cell* **13**, 67–76 (2004).
- Birse, C. E., Minvielle-Sebastia, L., Lee, B. A., Keller, W. & Proudfoot, N. J. Coupling termination of transcription to messenger RNA maturation in yeast. *Science* **280**, 298–301 (1998).
- Kim, M., Ahn, S. H., Krogan, N. J., Greenblatt, J. F. & Buratowski, S. Transitions in RNA polymerase II elongation complexes at the 3′ ends of genes. *EMBO J.* **23**, 354–364 (2004).
- Connelly, S. & Manley, J. L. A CCAAT box sequence in the adenovirus major late promoter functions as part of an RNA polymerase II termination signal. *Cell* **57**, 561–571 (1989).
- Proudfoot, N. J. How RNA polymerase II terminates transcription in higher eukaryotes. *Trends Biochem. Sci.* **14**, 105–110 (1989).
- Scholes, D. T., Banerjee, M., Bowen, B. & Curcio, M. J. Multiple regulators of Ty1 transposition in *Saccharomyces cerevisiae* have conserved roles in genome maintenance. *Genetics* **159**, 1449–1465 (2001).
- Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P. & Bork, P. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.* **12**, 47–56 (2002).
- Meinhart, A. & Cramer, P. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* **430**, 223–226 (2004).
- Tong, A. H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
- Stern, D. E., Lee, J. M., Hardin, S. E. & Greenleaf, A. L. The yeast carboxyl-terminal repeat domain kinase CTDK-I is a divergent cyclin-cyclin-dependent kinase complex. *Mol. Cell. Biol.* **15**, 5716–5724 (1995).
- Skaar, D. A. & Greenleaf, A. L. The RNA polymerase II CTD kinase CTDK-I affects pre-mRNA 3′ cleavage/polyadenylation through the processing component Pti1p. *Mol. Cell* **10**, 1429–1439 (2002).
- Dheur, S. A., Pti1p and Ref2p found in association with the mRNA 3′ end formation complex direct snoRNA maturation. *EMBO J.* **22**, 2831–2840 (2003).
- Nedea, E. et al. Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3′-ends. *J. Biol. Chem.* **278**, 33000–33010 (2003).
- Petfalski, E., Dandekar, T., Henry, Y. & Tollervey, D. Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. *Mol. Cell. Biol.* **18**, 1181–1189 (1998).
- Qu, L. H. et al. Seven novel methylation guide small nucleolar RNAs are processed from a common polycistronic transcript by Pti1p and RNase III in yeast. *Mol. Cell. Biol.* **19**, 1144–1158 (1999).
- Xue, Y. et al. *Saccharomyces cerevisiae* RAI1 (YGL246c) is homologous to human DOM3Z and encodes a protein that binds the nuclear exoribonuclease Rat1p. *Mol. Cell. Biol.* **20**, 4006–4015 (2000).
- Amberg, D. C., Goldstein, A. L. & Cole, C. N. Isolation and characterization of RAT1: an essential gene

- of *Saccharomyces cerevisiae* required for the efficient nucleocytoplasmic trafficking of mRNA. *Genes Dev.* **6**, 1173–1189 (1992).
19. Hammell, C. M. *et al.* Coupling of termination, 3' processing, and mRNA export. *Mol. Cell. Biol.* **22**, 6441–6457 (2002).
  20. Vasudevan, S. & Peltz, S. W. Nuclear mRNA surveillance. *Curr. Opin. Cell Biol.* **15**, 332–337 (2003).
  21. Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102**, 765–775 (2000).
  22. Das, B., Butler, J. S. & Sherman, F. Degradation of normal mRNA in the nucleus of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **23**, 5502–5515 (2003).
  23. Yonaha, M. & Proudfoot, N. J. Transcriptional termination and coupled polyadenylation *in vitro*. *EMBO J.* **19**, 3770–3777 (2000).
  24. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
  25. Solinger, J. A., Pascolini, D. & Heyer, W. D. Active-site mutations in the Xrn1p exoribonuclease of *Saccharomyces cerevisiae* reveal a specific role in meiosis. *Mol. Cell. Biol.* **19**, 5930–5942 (1999).
  26. Ujvari, A., Pal, M. & Luse, D. S. RNA polymerase II transcription complexes may become arrested if the nascent RNA is shortened to less than 50 nucleotides. *J. Biol. Chem.* **277**, 32527–32537 (2002).
  27. West, S., Gromak, N. & Proudfoot, N. Human 5' → 3' exonuclease XRN2 promotes transcription termination from co-transcriptional cleavage sites. *Nature* doi:10.1038/nature03035 (this issue).
  28. Krogan, N. J. *et al.* RNA polymerase II elongation factors of *Saccharomyces cerevisiae*: a targeted proteomics approach. *Mol. Cell. Biol.* **22**, 6979–6992 (2002).
  29. Zhao, J., Kessler, M., Helmling, S., O'Connor, J. P. & Moore, C. Pta1, a component of yeast CF II, is required for both cleavage and poly(A) addition of mRNA precursor. *Mol. Cell. Biol.* **19**, 7733–7740 (1999).
  30. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA* **94**, 13057–13062 (1997).

Supplementary Information accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank C. Moore and her laboratory for advice on polyadenylation reactions; N. Proudfoot, S. McCracken and B. Blencowe for sharing unpublished information; and the Taplin Mass Spectroscopy facility at Harvard Medical School and D. Richards of Affinium Pharmaceuticals for mass spectroscopy. N.J.K. was supported by a Doctoral Fellowship from the Canadian Institutes of Health Research (CIHR). O.J.R. is supported by funding from the Bauer Center for Genomics Research. This research was supported by grants to S.B. from the US National Institutes of Health, and to J.E.G. from the Canadian Institutes of Health Research, the Ontario Genomics Institute, and the National Cancer Institute of Canada with funds from the Canadian Cancer Society. L.V. is a Fellow and S.B. is a Scholar of the Leukemia and Lymphoma Society.

**Authors' contributions** All authors contributed to the conception and design of the experiments. M.K. performed the CTD affinity purification of Rtt103, RT-PCR analysis, ChIP analyses and Rat1 mutagenesis. N.K. did the Rtt103-TAP purification and the SGA analysis. L.V. conducted the immunoblotting, exonuclease and polyadenylation assays. E.N. performed the Pcf11 and Rna15 TAP purifications. The microarray experiments were done by M.K. and O.J.R. The paper was written by M.K., J.G. and S.B. with input from the other authors.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to S.B. (steveb@hms.harvard.edu).

## Human 5' → 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites

Steven West\*, Natalia Gromak\* & Nick J. Proudfoot

Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

\* These authors contributed equally to this work

Eukaryotic protein-encoding genes possess poly(A) signals that define the end of the messenger RNA and mediate downstream transcriptional termination by RNA polymerase II (Pol II)<sup>1</sup>. Termination could occur through an 'anti-termination' mechanism whereby elongation factors dissociate when the poly(A) signal is encountered, producing termination-competent Pol II<sup>2,3</sup>. An alternative 'torpedo' model postulated that poly(A) site cleavage provides an unprotected RNA 5' end that is degraded by 5' → 3' exonuclease activities (torpedoes) and so

induces dissociation of Pol II from the DNA template<sup>1,4</sup>. This model has been questioned because unprocessed transcripts read all the way to the site of transcriptional termination before upstream polyadenylation<sup>5–7</sup>. However, nascent transcripts located 1 kilobase downstream of the human  $\beta$ -globin gene poly(A) signal are associated with a co-transcriptional cleavage (CoTC) activity<sup>8</sup> that acts with the poly(A) signal to elicit efficient transcriptional termination. The CoTC sequence is an autocatalytic RNA structure that undergoes rapid self-cleavage<sup>9</sup>. Here we show that CoTC acts as a precursor to termination by presenting a free RNA 5' end that is recognized by the human 5' → 3' exonuclease Xrn2. Degradation of the downstream cleavage product by Xrn2 results in transcriptional termination, as envisaged in the torpedo model.

To confirm the requirement of a CoTC sequence for transcriptional termination we compared transcription of the human  $\beta$ -globin gene (WT), driven by a Tat-inducible HIV-1 promoter, with that of constructs either lacking CoTC ( $\Delta$ CoTC) or bearing a mutant poly(A) signal ( $\Delta$ p(A)). The plasmids were transfected into HeLa cells and their nascent transcripts analysed by nuclear run-on (NRO) analysis (Fig. 1a). In WT (top panel), termination occurs very efficiently within CoTC as indicated by low signals over probes A (downstream of the CoTC element) and U3 (upstream of the promoter). Termination efficiency is significantly reduced when CoTC is deleted, as indicated by the high signal over probes A and U3 (middle panel), and the mutant p(A) causes an even greater termination defect (bottom panel), as shown previously<sup>8</sup>. In the accompanying study<sup>9</sup> the minimal autocatalytic, or core, CoTC element is also shown to promote Pol II termination by using the same NRO assay. We conclude that efficient transcriptional termination of the human  $\beta$ -globin gene requires CoTC.

The possibility that autocatalytic CoTC provides an entry site for a 5' → 3' exonuclease was tested by the depletion of mRNAs encoding such activities using RNA interference technology<sup>10</sup>. The major nuclear 5' → 3' exonuclease in humans is Xrn2 (refs 11, 12), a 950-residue protein homologous to *Saccharomyces cerevisiae* Rat1p (ref. 13). Rat1p is known to be involved in 5.8S rRNA processing<sup>14</sup>, snoRNA maturation<sup>15</sup> and nuclear-cytoplasmic RNA transport<sup>13</sup>. Analyses by reverse transcriptase-mediated polymerase chain reaction (RT-PCR) and western blotting of Xrn2 mRNA and protein isolated from mock or Xrn2-specific short interfering RNA (siRNA)-treated HeLa cells are presented in Fig. 1b. An 80% decrease in the amount of Xrn2, but not actin, mRNA was observed. Similarly, western blotting with Xrn2 antibody<sup>12</sup> revealed that both forms of Xrn2 were significantly decreased (about fourfold), in comparison with transcription factor IIH (TFIIH). The smaller 70-kDa Xrn2 protein might result from the recognition of an internal methionine initiation codon<sup>16</sup>.

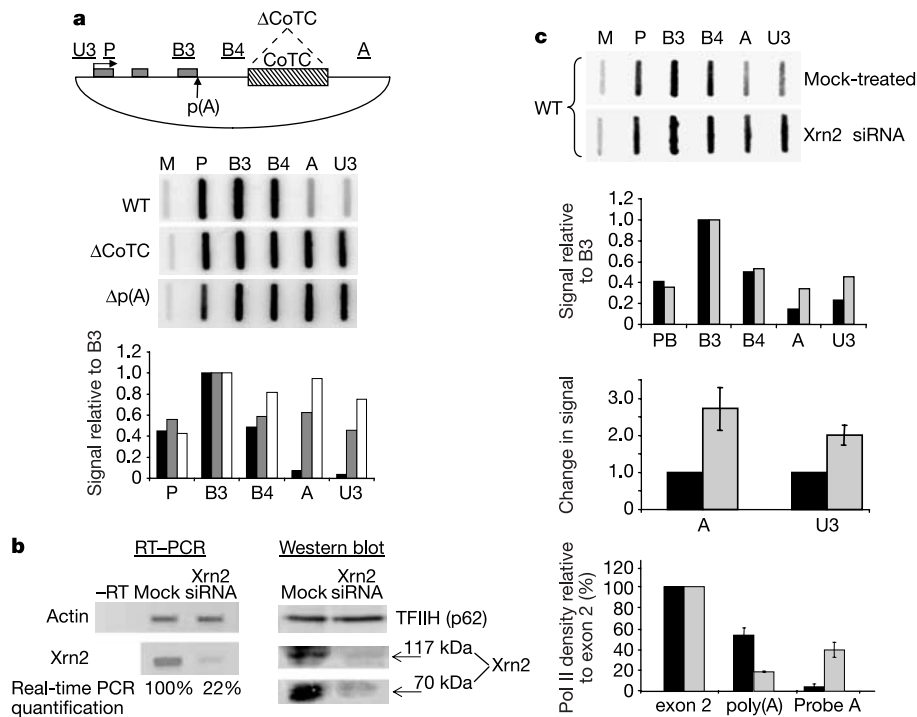
To test the effect of Xrn2 depletion on transcriptional termination we performed NRO on Xrn2 knock-down HeLa cells that were transiently transfected with the WT plasmid (Fig. 1c, top graph). A reproducible twofold to threefold increase in transcripts over probes A and U3 in comparison with mock-treated cells indicates a role for Xrn2 in transcriptional termination (Fig. 1c, middle graph). We believe that the observed effect might be an under-representation: some cells will remain termination-competent because Xrn2 knock-down is not 100% efficient. The same experiment was conducted with control (luciferase-specific) siRNA, resulting in an identical profile to that of the mock-treated cells (data not shown). Although the level of transcriptional termination with the WT plasmid was lower in the mock-treated than the untreated cells (compare Fig. 1c with Fig. 1a), we suspect that the multiple transfections necessary for successful RNA-mediated interference (RNAi) might be responsible for this. Taken together, these data show that depletion of Xrn2 results in a decrease in termination efficiency of the  $\beta$ -globin gene. To confirm these data independently, however, we performed chromatin immunoprecipitation (ChIP) analysis with a Pol II-

specific antibody (Fig. 1c, bottom graph). These data are consistent with the run-on analysis in showing loss of Pol II beyond the pA signal in the untreated HeLa cells transfected with the WT construct. CoTC-dependent termination is therefore associated with both a loss of nascent transcription and a release of Pol II from the chromatin template. Xrn2 depletion also shows a defect in termination efficiency as judged by the higher probe A signal. The lower Pol II signal over the pA region might be due to more efficient Pol II elongation across this sequence. The NRO and ChIP data above were confirmed by analysing steady-state nuclear RNA from transfected cells (Fig. 2). Transcript levels upstream and downstream of the CoTC element were measured using RNase protection analysis (RPA). If CoTC provides an entry site for a 5' → 3' exonuclease, we would predict that the transcript downstream of the cleavage site should be stabilized in Xrn2 knock-down HeLa cells. We compared the amounts of hybridization over the read-through probe A (downstream of CoTC) with the amounts of stable β-globin message detected by probe B3 (Fig. 2a). As indicated, the A/B3 ratio increased nearly fivefold after removal of the CoTC element in comparison with the WT construct (Fig. 2a, lanes 3 and 4), correlating with the NRO data (Fig. 1a). We also included the Δp(A) construct in this analysis and, as expected, obtained an even larger (sixfold) increase in read-through signal. Next we used the same probes to look at the stability of the downstream CoTC product after Xrn2 depletion (Fig. 2a, lanes 8 and 9). When the WT construct was transfected, Xrn2 depletion resulted in an up to fivefold increase in A-specific hybridization in comparison with the mock-treated cells. These data confirm the NRO results and strongly implicate Xrn2 activity in transcriptional termination as a result of degradation of the downstream CoTC product.

The possibility that Xrn2 gains access to the nascent RNA by

means of cleavage at the poly(A) site was also tested by probing steady-state transcripts from exon 3 up to the CoTC element (Fig. 2b). The strong p(A) band is the upstream product of poly(A) site cleavage and provides an indicator of the amount of β-globin mRNA. The band labelled 3' is the downstream product of poly(A) site cleavage and is of low intensity because it is not protected by a cap structure or a poly(A) tail. The intensity of this band does not vary on Xrn2 depletion. Therefore the 5' end generated by CoTC is the substrate for Xrn2 degradation and not poly(A) site cleavage. The intensity of the full-length, non-cleaved product band (NC) is unaltered by Xrn2 depletion, showing that the absence of Xrn2 has no effect on the poly(A) site cleavage mechanism. Also, when compared with the co-transfection Va control, the amount of p(A) RNA in the nucleus is unchanged after Xrn2 depletion (Fig. 2b). These data indicate that Xrn2-dependent termination is not an indirect consequence of depletion of polyadenylation factors that might have resulted from the retention of poly(A) RNA in the nucleus.

The possibility that any primary cleavage event in the 3' flanking region causes termination was tested by inserting a 52-nucleotide hammerhead ribozyme in place of the CoTC sequence<sup>17,18</sup> (to form plasmid pHH). The presence of the ribozyme did not cause transcriptional termination, as indicated by high signals over probes A and U3 in the NRO (compare Fig. 3a, top profile, with Fig. 1a). To check that the ribozyme was cleaving co-transcriptionally, we performed hybrid selection NRO<sup>7</sup> on HeLa cells transfected with pHH (Fig. 3a, middle and bottom profiles). Nascent transcripts continuous with the poly(A) site were selected with a biotinylated antisense RNA probe (positioned over the B4 region; see Fig. 1a). No such transcripts extended downstream of the hammerhead ribozyme, as indicated by the near-background amounts of signal



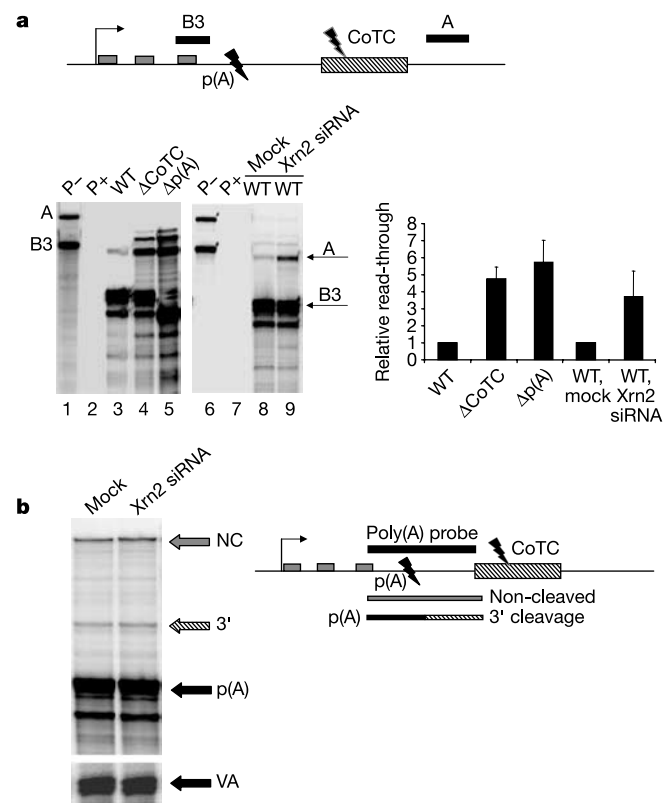
**Figure 1** RNAi-mediated depletion of human Xrn2 inhibits Pol II termination. **a**, Diagram of the human β-globin gene, with exons shown as grey boxes and CoTC element as a hatched box. NRO probe positions are underlined. The panels below show NRO signals for WT, ΔCoTC and Δp(A) transfections. M (empty M13 vector) shows the background signal. Signal quantification was corrected relative to B3 (taken as 1.0), as shown in the graph: black bars, WT; grey bars, ΔCoTC; white bars, Δp(A). **b**, Left, RT-PCR analysis of β-actin and Xrn2 mRNAs in mock and Xrn2 siRNA-treated cells. Right, western blot with Xrn2 antibody showing 117-kDa and 70-kDa Xrn2 variants in cells treated with Xrn2 siRNA

relative to constant TFIIH (p62). **c**, NRO analysis of WT transfected into mock-treated and Xrn2 siRNA-treated cells. Quantification is shown in the top graph for WT reporter (black bars, mock; grey bars, Xrn2 siRNA). Middle graph: change in signal over read-through probes A and U3 between mock-treated (black bars) and Xrn2-specific siRNA-treated (grey bars) cells. Bottom graph: Pol II ChIP analysis across the β-globin gene using primers for exon 2, polyA and probe A regions. Pol II density is shown as enrichment over the background, with exon 2 taken as 100%. Black bars, untreated; grey bars, Xrn2 siRNA. Error bars (s.d.) throughout are based on at least three experiments.



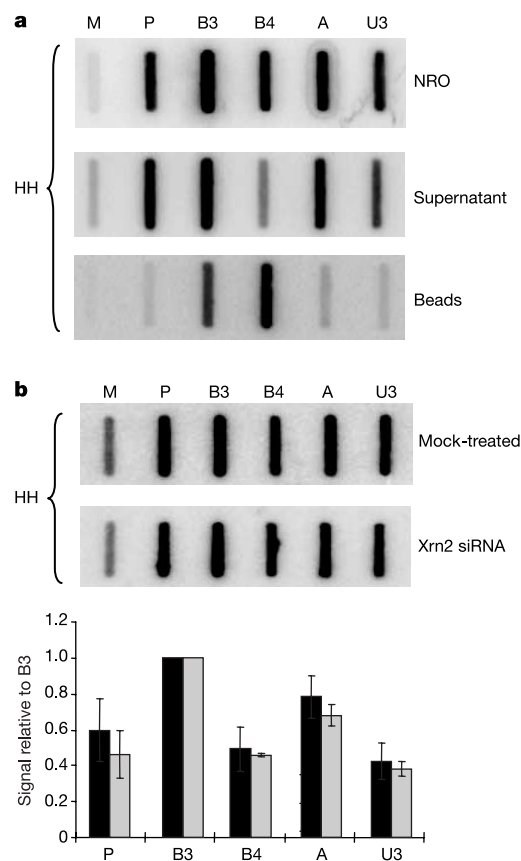
over probes A and U3 (Fig. 3a, bottom profile), confirming that the ribozyme cleaves co-transcriptionally. Note that probe B3, which is upstream of the selection probe, shows a hybridization signal because Pol II transcribes into the selected RNA region during the NRO transcription reaction. The downstream products of ribozyme cleavage are present in the non-selected, supernatant fraction (note the large amounts of signal over probes A and U3 in Fig. 3a, middle profile). This experiment shows that efficient transcriptional termination requires autocatalytic CoTC cleavage and cannot occur through other forms of co-transcriptional cleavage. We also investigated the effect of Xrn2 knock-down on pHH transcription. As shown in Fig. 3b, the NRO profile was unchanged when pHH was transfected into Xrn2 knock-down HeLa cells, indicating that any effect of Xrn2 on human  $\beta$ -globin transcriptional termination is specific to CoTC (Fig. 2c). These data indicate that hammerhead ribozyme cleavage does not provide the necessary substrate for exonuclease degradation. This might be explained by the fact that hammerhead ribozyme cleavage results in a 5' hydroxy terminus<sup>19</sup>, whereas CoTC generates a 5' phosphate terminus<sup>9</sup>. Previously, 5'  $\rightarrow$  3' exonucleases such as Xrn2 have been shown to require a 5' phosphate for activity<sup>20</sup>. Alternatively Xrn2 may display a sequence or structural preference for CoTC-generated 5' termini.

The efficient production of mRNA depends on coupled transcription and mRNA processing<sup>3</sup>. This is clearly evident from the combined requirement for Pol II termination of CoTC in the 3'

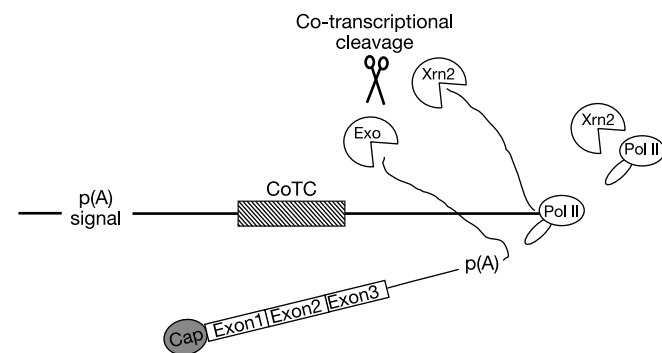


**Figure 2** RNAi-mediated depletion of human Xrn2 protein increases steady-state read-through transcription. **a**, RPA of WT,  $\Delta$ CoTC and  $\Delta$ p(A) transfections (lanes 3–5) and WT transfected into mock-treated and Xrn2 siRNA-treated cells (lanes 8 and 9). Lanes P + and P – are RNA probes alone incubated with or without the RNase treatment (including carrier tRNA). B3 band in lane 5 is slightly smaller because of a mismatch between the p(A) signal mutation and probe. The diagram at the top depicts two riboprobes, B3 and A. Jagged arrows denote RNA cleavage at the poly(A) site (p(A)) and CoTC. The graph shows quantification with error bars (s.d.) indicated. **b**, RPA of WT transfected into mock-treated and Xrn2 siRNA-treated cells. The diagram at the right shows the probe position. The VA co-transfection control signal is also shown. A-specific as well as the 3'-cleaved and non-cleaved products were not detected in the cytoplasm (data not shown).

flanking region and upstream cleavage and polyadenylation. Our new data support a revised torpedo model (Fig. 4) in which rapid autocatalytic CoTC results in cleavage of the nascent transcript close to the site of Pol II termination, allowing time for Xrn2 to strip off the residual nascent transcript and thereby promote termination. At the same time a 3'  $\rightarrow$  5' exonuclease activity as part of the exosome may quickly degrade the transcript back to the poly(A) signal<sup>21</sup>. The



**Figure 3** Hammerhead ribozyme cleavage does not promote Xrn2-dependent termination. **a**, NRO (top) and hybrid selection (middle and bottom) of pHH containing hammerhead ribozyme in place of the CoTC element. Top: NRO profile of the pHH construct. Middle: transcripts not selected by biotinylated  $\beta$ 4 probe. Bottom: transcripts selected by biotinylated  $\beta$ 4 probe. **b**, NRO analysis of pHH transfected into mock-treated cells (black bars) or cells treated with Xrn2 siRNA (grey bars). Corrected hybridization signals are shown in the graph below with error bars (s.d.) indicated.



**Figure 4** Model for Pol II transcriptional termination. Autocatalytic CoTC cleavage acts as a precursor to termination by presenting a free 5' end that is recognized by the 5'  $\rightarrow$  3' exonuclease Xrn2. Subsequent degradation of this downstream cleavage product by Xrn2 leads to Pol II transcriptional termination analogous to the original torpedo model<sup>1,4</sup>. The 3' end of the cleaved RNA is likely to be degraded by the exosome (Exo)<sup>21</sup>.

observation that CoTC activity is in part autocatalytic<sup>9</sup> suggests that evolution has generated a built-in mechanism to ensure that transcriptional termination occurs efficiently. Recent studies in *S. cerevisiae* have independently demonstrated a role for the Xrn2 homologue, Rat1p, in Pol II termination<sup>22</sup>. This points to the generality of 5' → 3' exonuclease, torpedo-dependent termination. □

## Methods

### Plasmid construction

WT, ΔCoTC and Δp(A) constructs correspond to βΔ5-7, βΔ5-10 and βΔ5-7p(A)mut plasmids<sup>5</sup>. The Tat plasmid has been previously described<sup>23</sup>. pHH construct was produced by the insertion of duplex oligonucleotides HHF (5'-CCTGTCACCGGATGTGTTTCCGGTCTGATGAGTCCGTGAGGACGAAACG-3') and HHR (5'-CCTGTTTCGTCTCTACCGACTCATCAGACCGAAAACACATCCGGTGACG-3')<sup>17</sup> into a vector prepared by PCR from βΔ5-7 by using Pfu Turbo DNA polymerase (Stratagene) with the VF (5'-CAGGAACTATTACTCAAGGTA-3') and VR (5'-TTGAATCCTTTCTGAGGGATG-3') oligonucleotide pair. pBex3 for B3 riboprobe synthesis was prepared by ligation of the 63,424–63,708 (Genbank accession number U01317) fragment into pGEM4 vector (Promega) cut with *HincII*. A-specific riboprobe was generated from pGEM vector containing NRO probe A<sup>7</sup>. The poly(A) riboprobe was obtained from a plasmid made by blunt-end ligation of the *AvrII/BstXI* fragment of βΔ5-7 into pGEM4 (Promega) cut with *HincII*. The plasmid from which the Va riboprobe was generated has been described previously<sup>7</sup>.

### NRO and hybrid selection probes

β-Globin gene single-stranded DNA NRO probes as well as P and U3 have been described previously<sup>7,24</sup>. A-single-stranded DNA NRO probe was prepared by insertion of an *AatII/XmnI* digestion of βΔ5-7 into M13mp19 vector (New England Biolabs) digested with *HincII*. The β4 hybrid selection probe has been described previously<sup>7</sup>.

### RPA probes

pBex3 was linearized with *EcoRI* and Va, poly(A) and A were linearized with *HindIII*. Each template (0.5 μg) was transcribed with T7 RNA polymerase (Promega) for Va, p(A) and A, or SP6 RNA polymerase (Roche) for pBex3. Each transcription reaction was performed in accordance to the manufacturer's guidelines and supplemented with 20 μCi of [ $\alpha$ -<sup>32</sup>P]UTP. Full-length riboprobe was gel-purified using standard procedures.

### Transfection procedure

Sub-confluent HeLa cells were transiently transfected with 20 μg of test plasmid and 3 μg of the Tat plasmid by using Lipofectamine 2000 (Invitrogen) in accordance with the manufacturer's guidelines. For RNAi treatment, two siRNA (or buffers alone in mock treatment) transfections were performed before plasmid transfection<sup>25,26</sup>.

### RNA interference

siRNA (Dharmacon) mRNA targets were for Xrn2 5'-AAGAGUACAGAUGAUGUUU-3', and for luciferase 5'-CGTACGCGGAATACTTCA-3'. Oligonucleotides for Xrn2 RT-PCR analysis were 5'-TCCTTCGGCTTAATGTTCTTC-3' and 5'-AGATGTGAAACTCGTATTAGG-3'. PCR cycles were performed at 94 °C for 30 s, 52 °C for 30 s and 72 °C for 30 s. For β-actin message analysis RT-PCR primer sequences were 5'-CGTGATGGTGGGCATGGGTCAG-3' and 5'-CTTAATGTACGCACGATTTC-3', and PCR cycles were performed at 94 °C for 30 s, 62 °C for 30 s and 72 °C for 30 s. For Xrn2 and β-actin, cycle numbers were restricted to maintain amplification in the linear range. The same primers were employed for real-time PCR quantification.

### Western blotting

Western blotting was performed by standard immunoblotting procedure with ECF western blotting kit (Amersham Biosciences). Antibodies for western blot analysis were rabbit anti-Xrn2 antibody<sup>12</sup>, rabbit anti-TFIIF (p62) (Q-19) (Santa Cruz Biotechnology), anti-rabbit fluorescein-conjugated antibody and anti-fluorescein alkaline phosphatase conjugate (Amersham Biosciences). The signals were obtained with Molecular Dynamics Storm software.

### ChIP analysis

ChIP analysis of the transfected plasmids was carried out as described previously in the Upstate Biotechnology (Lake Placid, NY) protocol and in ref. 27. This technique is difficult to perform with transiently transfected gene constructs because the large amount of untranscribed plasmid trapped at the nuclear membrane gives a high background ChIP signal. However, extensive washing of the nuclear fraction allows significant amounts of globin DNA to be selected by immunoprecipitation and detected by real-time PCR. Pol II antibody (H-224) was obtained from Santa Cruz Biotechnology. Immunoprecipitated, non-precipitated and input DNA were used as templates for real-time quantitative PCR, performed using a Corbett Research Rotor-Gene GG-3000 machine. The PCR mixture contained QuantiTect SYBR Green PCR master mix (Qiagen), 1 μl of the template DNA and the following primers: ex2(F), 5'-TTGGACCCAGAGGTTCTTTG-3'; ex2(R), 5'-GAGCCAGGCCATCACTAAAG-3'; probe A(F), 5'-TTGCCTTCTCTGTTTGTGCTC-3'; probe A(R), 5'-CCGTGTTGAGATCCAGTTC-3'; polyA(F), 5'-AAAAGGGAATGTGGGAGGTC-3'; and polyA(R), 5'-AGCCTCACCTTCTTTCATGG-3'. Cycling parameters were 95 °C for 15 min, followed by 45 cycles of 95 °C for 15 s, 57.5 °C for 30 s

and 72 °C for 30 s. Fluorescence intensities were plotted against the number of cycles by using an algorithm provided by the manufacturer.

### RNA analysis

The methods used for NRO and RPA have been described previously<sup>7,24</sup>. Relative ratios of the nascent and steady-state RNA amounts were determined with Molecular Dynamics ImageQuant software (version 1.11). Hybrid selection NRO performed on pHH construct was performed as previously described<sup>7</sup>; however, to avoid hammerhead ribozyme cleavage occurring during RNA isolation, TRIzol (Invitrogen) denaturation protocol, as well as buffers lacking Mg<sup>2+</sup> (required for the ribozyme cleavage), was employed before selection.

Received 28 June; accepted 16 September 2004; doi:10.1038/nature03035.

- Proudfoot, N. J. How RNA polymerase II terminates transcription in higher eukaryotes. *Trends Biochem. Sci.* **14**, 105–110 (1989).
- Logan, J., Falk-Pedersen, E., Darnell, J. E. Jr & Shenk, T. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc. Natl Acad. Sci. USA* **84**, 8306–8310 (1987).
- Proudfoot, N. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr. Opin. Cell Biol.* **16**, 272–278 (2004).
- Connelly, S. & Manley, J. L. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.* **2**, 440–452 (1988).
- Osheim, Y. N., Proudfoot, N. J. & Beyer, A. L. EM visualization of transcription by RNA polymerase II: downstream termination requires a poly(A) signal but not transcript cleavage. *Mol. Cell* **3**, 379–387 (1999).
- Bauren, G., Belikov, S. & Wieslander, L. Transcriptional termination in the Balbiani ring 1 gene is closely coupled to 3'-end formation and excision of the 3'-terminal intron. *Genes Dev.* **12**, 2759–2769 (1998).
- Dye, M. J. & Proudfoot, N. J. Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. *Mol. Cell* **3**, 371–378 (1999).
- Dye, M. J. & Proudfoot, N. J. Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. *Cell* **105**, 669–681 (2001).
- Teixeira, A. *et al.* Auto-catalytic RNA cleavage in the human β-globin pre-mRNA promotes transcription termination. *Nature* doi:10.1038/nature03032 (this issue).
- Cerutti, H. RNA interference: travelling in the cell and gaining functions. *Trends Genet.* **19**, 39–46 (2003).
- Zhang, M. *et al.* Cloning and mapping of the *XRN2* gene to human chromosome 20p11.1–p11.2. *Genomics* **59**, 252–254 (1999).
- Lejeune, F., Li, X. & Maquat, L. E. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylation, and exonucleolytic activities. *Mol. Cell* **12**, 675–687 (2003).
- Amberg, D. C., Goldstein, A. L. & Cole, C. N. Isolation and characterization of *RAT1*: an essential gene of *Saccharomyces cerevisiae* required for the efficient nucleocytoplasmic trafficking of mRNA. *Genes Dev.* **6**, 1173–1189 (1992).
- Henry, Y. *et al.* The 5' end of yeast 5.8S rRNA is generated by exonucleases from an upstream cleavage site. *EMBO J.* **13**, 2452–2463 (1994).
- Petfalski, E., Dandekar, T., Henry, Y. & Tollervey, D. Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. *Mol. Cell Biol.* **18**, 1181–1189 (1998).
- Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99**, 16899–16903 (2002).
- Samarsky, D. A. *et al.* A small nucleolar RNA:ribozyme hybrid cleaves a nucleolar RNA target *in vivo* with near-perfect efficiency. *Proc. Natl Acad. Sci. USA* **96**, 6609–6614 (1999).
- Huertas, P. & Aguilera, A. Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell* **12**, 711–721 (2003).
- Hammann, C. & Lilley, D. M. Folding and activity of the hammerhead ribozyme. *ChemBiochem* **3**, 690–700 (2002).
- Stevens, A. & Poole, T. L. 5' exonuclease 2 of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **270**, 16063–16069 (1995).
- Torchet, C. *et al.* Processing of 3'-extended read-through transcripts by the exosome can generate functional mRNAs. *Mol. Cell* **9**, 1285–1296 (2002).
- Kim, M. *et al.* The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* doi:10.1038/nature03041 (this issue).
- Adams, S. E. *et al.* Synthesis of a gene for the HIV transactivator protein TAT by a novel single-stranded approach involving *in vivo* gap repair. *Nucleic Acids Res.* **16**, 4287–4298 (1988).
- Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
- Wagner, E. J. & Garcia-Blanco, M. A. RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol. Cell* **10**, 943–949 (2002).
- Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A. & Smith, C. W. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell* **13**, 91–100 (2004).
- Kadener, S., Fededa, J. P., Rosbash, M. & Kornbliht, A. R. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc. Natl Acad. Sci. USA* **99**, 8185–8190 (2002).

**Acknowledgements** We thank M. Dye for expert advice on NRO analysis, M. Wollerton for advice on RNAi, I. Martianov for help with the western blot, L. Maquat and F. Lejeune for providing Xrn2 antibody, and A. Teixeira, A. Akoulitchev and members of N.J.P.'s laboratory for advice and encouragement. This work was supported by a Programme Grant from the Wellcome Trust.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to N.J.P. (nicholas.proudfoot@path.ox.ac.uk).

# Autocatalytic RNA cleavage in the human $\beta$ -globin pre-mRNA promotes transcription termination

Alexandre Teixeira<sup>1</sup>, Abdessamad Tahiri-Alaoui<sup>1</sup>, Steve West<sup>1</sup>, Benjamin Thomas<sup>1</sup>, Aroul Ramadass<sup>2</sup>, Igor Martianov<sup>1</sup>, Mick Dye<sup>1</sup>, William James<sup>1</sup>, Nick J. Proudfoot<sup>1</sup> & Alexandre Akoulitchev<sup>1</sup>

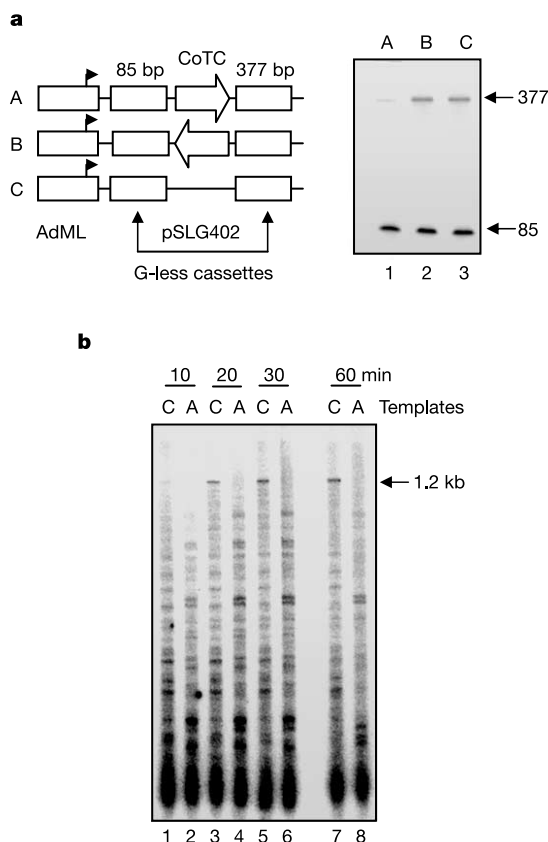
<sup>1</sup>Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

<sup>2</sup>The Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK

New evidence indicates that termination of transcription is an important regulatory step, closely related to transcriptional interference<sup>1</sup> and even transcriptional initiation<sup>2</sup>. However, how this occurs is poorly understood. Recently, *in vivo* analysis of transcriptional termination for the human  $\beta$ -globin gene revealed a new phenomenon—co-transcriptional cleavage (CoTC)<sup>3</sup>. This primary cleavage event within  $\beta$ -globin pre-messenger RNA, downstream of the poly(A) site, is critical for efficient transcriptional termination by RNA polymerase II<sup>3</sup>. Here we show that the CoTC process in the human  $\beta$ -globin gene involves an RNA self-cleaving activity. We characterize the

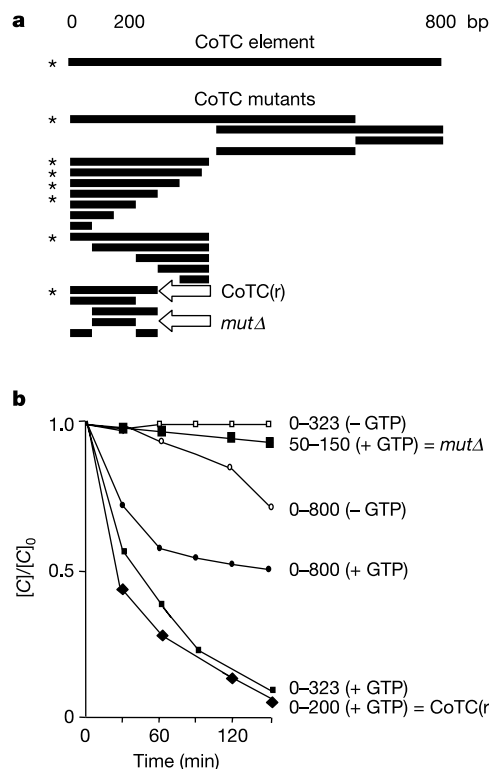
autocatalytic core of the CoTC ribozyme and show its functional role in efficient termination *in vivo*. The identified core CoTC is highly conserved in the 3' flanking regions of other primate  $\beta$ -globin genes. Functionally, it resembles the 3' processive, self-cleaving ribozymes described for the protein-encoding genes from the myxomycetes *Didymium iridis* and *Physarum polycephalum*, indicating evolutionary conservation of this molecular process. We predict that regulated autocatalytic cleavage elements within pre-mRNAs may be a general phenomenon and that functionally it may provide the entry point for exonucleases involved in mRNA maturation, turnover and, in particular, transcriptional termination.

To characterize the mechanism of co-transcriptional cleavage and its role in transcriptional termination, we first analysed the effect of the CoTC element *in vitro* in a reconstituted transcription assay with RNA polymerase II<sup>4</sup>. The CoTC element from the  $\beta$ -globin gene<sup>3</sup> was cloned in both orientations between two G-less cassettes (Fig. 1a) in a template designed to monitor transcription elongation, pSLG402 (ref. 5). pSLG402 itself was also used as a control for efficient transcriptional elongation, in accordance with its original design<sup>5</sup>. In the presence of the correctly oriented CoTC element, transcripts from the distal G-less cassette were hardly detectable after the treatment with T1 RNase (Fig. 1a, lane A). To characterize the effect further we also compared the full-length run-off transcripts without RNase treatment. The time course demonstrated that full-length transcripts of 1.2 kilobases gradually accumulated from the control template; however, the termination template failed to produce detectable full-length transcripts



**Figure 1** Effect of the CoTC element on reconstituted transcription *in vitro*.

**a**, Reconstituted transcription from termination (A, lane 1) and control (B, C, lanes 2 and 3, respectively) templates after treatment with T1 RNase. Arrows mark the positions of the G-less transcripts from the promoter proximal (85 nt) and distal (377 nt) cassettes. **b**, Time course of reconstituted run-off transcription from the termination (A, even-numbered lanes) and control (C, odd-numbered lanes) templates. The arrow marks the position of the full-length transcript (1,200 nt). Abbreviation: kb, kilobases.



**Figure 2** Mapping of the minimal catalytic core within the CoTC element. **a**, Deletion analysis of the synthetic CoTC transcripts. Deletion mutants of the synthetic CoTC RNA were tested for GTP-dependent degradation in a protein-free environment. Transcripts that displayed catalytic activity are marked with asterisks; arrows mark the minimal ribozyme core of the CoTC element (1–200), designated CoTC(r), and the deletion mutant (50–150), designated *mutΔ*. **b**, Kinetics of GTP-dependent degradation for the deletion mutants.



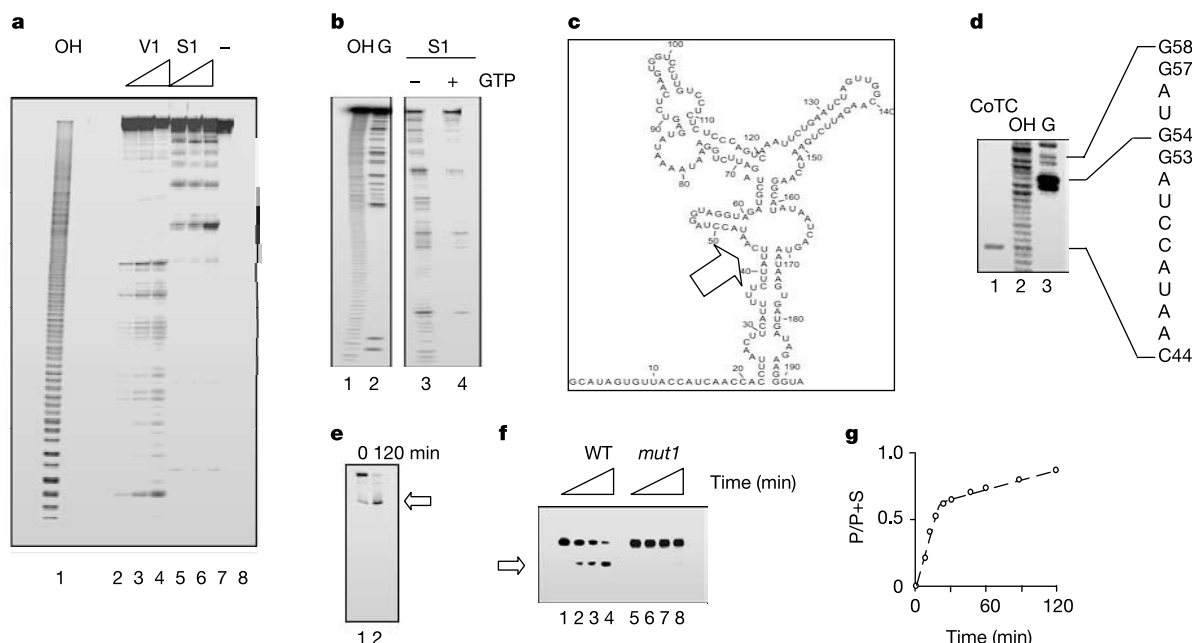
(Fig. 1b, compare A with C). Moreover, we observed additional partial-length products that had low stability after the first 30 min. We concluded that under conditions of reconstituted transcription, the presence of the CoTC element in the correct orientation prevented the production of stable transcripts from the distal portion of the template. This could be attributed to several causes, ranging from transcriptional pausing or termination to the short half-life of the CoTC transcript.

To analyse further the properties of transcripts containing the CoTC element, we decided to uncouple transcription from the subsequent RNA processing steps. We synthesized  $^{32}\text{P}$ -labelled synthetic CoTC RNA by using T7 RNA polymerase as a potential substrate for specific RNA cleavage activities. Surprisingly, our analysis revealed that, unlike the control, the CoTC RNA had a short half-life with a distinct pattern of degradation/cleavage products in the presence of the nucleophilic cofactors  $\text{Mg}^{2+}$  and GTP<sup>6</sup> (Supplementary Fig. 1a). These products appeared under protein-free conditions, because the cleavage activity was maintained after gel purification in the presence of 8 M urea, after tryptic digestion, after extraction with phenol/chloroform, and in the presence of SDS<sup>7</sup> (Supplementary Fig. 4a). The degradation kinetics of the full-length synthetic CoTC RNA corresponds to a first-order catalytic reaction (Supplementary Fig. 1b) with  $k = 0.015 \text{ min}^{-1}$  and a half-life of 38 min, which is longer than for some classes of self-cleaving ribozymes<sup>8</sup>. However, the recently described autocleaving prokaryotic ribozyme in the *Bacillus subtilis* *glmS* mRNA<sup>9</sup> has a half-life of 4 h ( $k = 0.003 \text{ min}^{-1}$ ) under certain conditions. This is still  $10^3$ -fold faster than spontaneous cleavage of an unconstrained RNA linkage<sup>10</sup> and tenfold faster than spontaneous nucleophilic cleavage of a constrained RNA linkage<sup>11</sup>. Our results indicate that CoTC RNA undergoes a single-molecule autocatalytic cleavage reaction, as described for certain classes of RNA

ribozyme *in vivo* and *in vitro*<sup>6,8</sup>. The rates of autocatalytic cleavage can be subject to allosteric regulation. For example, *glmS* mRNA cleavage is activated almost  $10^3$ -fold, with  $k_{\text{max}} = 1 \text{ min}^{-1}$ , by its effector GlcN6P (ref. 9). If the complex kinetics of CoTC cleavage is taken into account (see below), it might well be subject to regulation and reach higher rates of cleavage *in vivo*. In fact, preliminary biochemical analysis confirms the existence of regulatory effectors and RNA chaperones for the CoTC element (A.A., unpublished data).

Next, we mapped the minimal region sufficient for autocatalytic cleavage of the synthetic CoTC RNA. We generated a series of 5'- and 3'-end deletion mutants spanning the 850 nucleotides (nt) of the CoTC element (Fig. 2a). In the cleavage analysis we monitored two parameters: the half-life of the transcripts and the dependence on GTP. A 200-nt region close to the 5' end of the CoTC element displayed a cleavage half-life of 15 min in the presence of GTP cofactor, with further deletions from the 5' and 3' ends (*mutΔ*) rendering the transcripts stable under the same conditions (Fig. 2b). We conclude that the ribozyme core, CoTC(r), is located within these 200 nt.

To establish more comprehensive evidence for the autocatalytic activity of CoTC(r), we investigated its structural properties in more detail. First, we analysed single-stranded and double-stranded secondary structures within CoTC(r) by using the nucleases S1 and T1, which are specific for single-stranded RNA, as well as the double-strand-specific nuclease V1. As shown in Fig. 3a, digestion with V1 and S1 identifies a series of single-stranded and double-stranded regions (for a detailed analysis see Supplementary Fig. 2). We also analysed the secondary structure of CoTC(r) using a chemical modification approach (Supplementary Fig. 3). Single-stranded nucleosides were modified with dimethylsulphate (DMS) at A(N1), C(N3) and with 1-cyclohexyl-3-(2-morpholinoethyl)car-



**Figure 3** Analysis of the CoTC core secondary structure and mapping of the autocatalytic cleavage site. **a**, Cleavage profile for the 5'- $^{32}\text{P}$  CoTC(r) RNA (lane 8) treated with endonucleases V1 (lanes 2–4) and S1 (lanes 5–7). **b**, S1 cleavage profile of the CoTC(r) RNA (lane 3) in the absence or presence of 1 mM GTP (lanes 3 and 4). **c**, Secondary structure of the catalytic CoTC core, as defined by the folding algorithm<sup>13</sup> and the data from enzymatic and chemical analyses. The arrow marks the position of the primary autocatalytic cleavage site at C44. **d**, The product of the autocatalytic cleavage isolated and resolved on a sequencing gel. OH identifies the alkaline hydrolysis ladder marking

each nucleoside position on the gel; G identifies guanosine positions after partial digestion of the CoTC(r) with T1. **e**, The CoTC element (1–323) is cleaved (lane 2, arrow) under protein-free conditions in the presence of GTP. **f**, CoTC(r), the minimal catalytic core, is cleaved in a time course conducted under protein-free conditions (lanes 1–4 correspond to 0, 5, 15 and 120 min, respectively). Cleavage was performed with the 5'- $^{32}\text{P}$  core RNA in the presence of 200 mM KCl. The arrow marks the primary cleavage site. *mut1* denotes a double mutant (C45C46) at the cleavage site. **g**, Profile of the biphasic product accumulation in the CoTC cleavage reaction. P, product; S, substrate.

bodiimide metho-*p*-toluene sulphonate (CMCT) at G(N1), U(N3). The positions of modified nucleosides were identified by primer extension with reverse transcriptase (Supplementary Fig. 3).

The data from the analyses of enzymatic and chemical modification were used to constrain simulations of secondary structure with the Mfold 3.1 program<sup>12</sup>. The resulting structure (Fig. 3c) is predicted to have a free energy change on folding of  $-32.87 \text{ kcal mol}^{-1}$  and has clear structural and sequence characteristics. One notable feature is the presence of an AU-rich stem at the base of the structure, formed by the 5' and 3' ends of the sequence. Earlier deletion analysis with *mutΔ* identified this AU-rich stem as essential for GTP-dependent cleavage (Fig. 2). The highly structured organization of CoTC(r) compares favourably in complexity of folding with riboswitches described previously<sup>13</sup>.

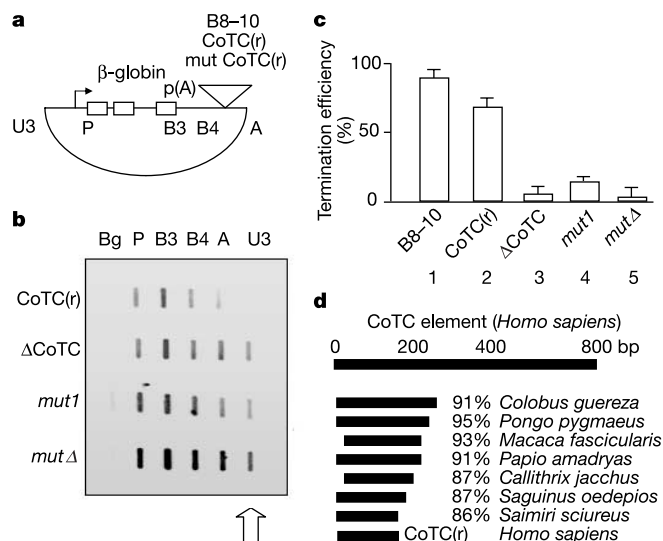
We also analysed the effect of GTP on the secondary structure of CoTC(r) (Fig. 3b). S1 digestion shows a loss of single-strand endoRNase cleavages in several positions in the presence of GTP (Fig. 3b), indicating significant changes in the secondary structure of CoTC(r). Classical studies of the group I ribozyme showed that nucleophilic attack by GTP was dependent on 2'- and 3'-hydroxy groups<sup>14,15</sup>. The diphosphate and monophosphate derivatives (GDP

and GMP) still supported the cleavage reaction, whereas treatment with dideoxy-GTP (ddGTP) resulted in a loss of activity. For CoTC(r) we also observed a loss of cleavage with ddGTP (Supplementary Fig. 4a), whereas GTP, GDP and GMP were equally effective cofactors. We did not detect any covalent association of GTP with the cleavage products (data not shown). We observed a high affinity of GTP for the structured CoTC RNA. This was shown by pull-down experiments with the <sup>32</sup>P-labelled CoTC RNA in the presence of a tRNA competitor with agarose beads cross-linked to either GTP or UTP. Only with GTP was the labelled RNA retained on the agarose beads (Supplementary Fig. 4c). Further analysis of the cleavage mechanism will delineate the exact role played by GTP.

Degradation profiles of CoTC(r) indicate the presence of a primary cleavage site (Fig. 3c–e), which is also resistant to SDS treatment<sup>7</sup> (Supplementary Fig. 4a). We identified the position of the site by isolating the cleavage product and resolving it on the sequencing gel against defined RNA markers (Fig. 3c, d). We then generated several double-point mutants around the cleavage site, as well as at the stem of the CoTC structure. In particular, the cleavage site mutant *mut1* (C45C46) was severely impaired in its function (Fig. 3f). Several additional mutants within the stem (C39C40, C40C41) also lost autocatalytic activity (data not shown). These mutant results are consistent with a functional role of the stem, as shown in the earlier degradation assay of *mutΔ* CoTC RNA (Fig. 2). We also confirmed that the 3' ends of the CoTC primary cleavage site could be labelled with pCp in a T4 RNA ligase reaction, whereas the 5' ends could be labelled with polynucleotide kinase after treatment with phosphatase (data not shown). This labelling specificity suggests that the cleavage reaction generates 3'-hydroxy and 5'-phosphate ends.

An analysis of product accumulation during CoTC(r) cleavage reveals biphasic kinetics (Fig. 3g)<sup>16</sup>. Here, the early cleavage of a portion of the substrate is observed at a high *k* of about  $1 \text{ min}^{-1}$  in the initial linear range, followed by slower cleavage of the remaining substrate. The explanation for the biphasic kinetics might lie with the alternative conformation of E–S (ribozyme–substrate) forms and slow exchange with the active conformation. Indeed, correct folding before cleavage analysis proved important during experimental analysis of CoTC(r). In the context of highly regulated transcriptional termination *in vivo*, the biphasic behaviour of CoTC(r) might accommodate multilayered regulation by effectors and RNA chaperones, as mentioned above. We are currently investigating further details of this autocatalytic reaction because we hope to be able to classify and relate the CoTC(r) element to other known classes of catalytic RNA<sup>6</sup>.

To confirm the biological relevance of the *in vitro* autocatalytic activity of CoTC(r) we determined whether the catalytic core of the CoTC element is implicated in the termination of RNA polymerase II-mediated transcription. The plasmid containing the β-globin gene was transfected into HeLa cells, followed by nuclear run-on (NRO) analysis of termination as described previously<sup>3</sup>. The β-globin gene contained one copy of the 200-nt core CoTC(r) or its mutant forms replacing the whole 3' flanking region (region B5–B10)<sup>3</sup>, as indicated. In the absence of CoTC(r), transcriptional termination is defective and NRO signals are detected far downstream of the cleavage site by probe U3 (Fig. 4a–c). After insertion of the minimal ribozyme core CoTC(r), more than 60% of polymerases successfully terminate (Fig. 4c). The cleavage-site mutant *mut1*, which is defective in catalytic cleavage (Fig. 3f), has a low termination efficiency (Fig. 4b, c). In addition, deletion mutant *mutΔ* (see Fig. 2b), as well as point mutants within the stem structure defective in cleavage (*mut 5* (C39C40) and *mut 6* (C40C41)), were similarly deficient in termination (data not shown). These experiments *in vivo* indicate strongly that the catalytic activity of the identified minimal ribozyme core of the CoTC element is essential in the transcriptional termination process.



**Figure 4** The catalytic ribozyme core of CoTC mediates termination *in vivo* and is conserved through evolution. **a**, Read-through *in vivo* assay: the 3' end of the β-globin gene (B5–10) was removed and replaced with the partial sequence B8–B10, or the CoTC core constructs (CoTC(r)) and its mutants, as described in the text. The density of the RNA polymerase II that escaped termination was monitored by run-on assay. **b**, The upper panel marked CoTC(r) of the read-through run-on assay shows the distribution of the RNA polymerase II density on the plasmid where the 3' end of the β-globin gene was replaced with CoTC(r). The lower panel marked ΔCoTC shows the result for the β-globin gene lacking the 3' end. The arrow marks the U3 probe for the polymerase that escaped termination. *mut1* denotes the mutant at the ribozyme cleavage site (C45C46), deficient in autocatalytic cleavage, as shown in Fig. 3. *mutΔ* denotes the deletion mutant 50–150 of the CoTC (Fig. 2), also deficient in autocatalytic cleavage. **c**, Efficiency of termination as measured by the run-ons for the read-through assay. 100% corresponds to the level of termination observed with the U3 probe from the β-globin gene containing its full 3' end, as described in ref. 3. The 3' end was replaced with the β-globin gene 3' end (B8–10) fragment containing the full CoTC element, the core ribozyme CoTC element, no core CoTC, cleavage mutant *mut1* or the deletion mutant *mutΔ*, respectively. Results are means ± s.d. for four duplicate sets of independent assays. The results are statistically significant ( $P < 0.05$ ). **d**, Conservation of the CoTC element in the non-coding 3' end of the homologues of the β-globin gene among primates. The diagram shows regions of high homology and percentage of identity between human CoTC and the primate genes. *S. sciureus* features the catalytic core of the CoTC element, CoTC(r), as the only highly preserved part of the sequence.

We finally turned our attention to the possibility that autocatalytic elements such as core CoTC might have a more general function in transcriptional termination. A limited search in the close taxonomy group of primates, in which sequence similarity would be the highest, reveals that  $\beta$ -globin genes from the available primate sequences contain in their 3' flanking regions sequences homologous to the full-length human CoTC element (Fig. 4d). The most highly conserved region corresponds exactly to the catalytic core CoTC(r). Because the 3' flanking region sequences are normally highly divergent, this observed homology is especially significant. Indeed, in *Saimiri sciureus* the highly conserved homologue of the ribozyme core CoTC was identified 1,456 nt downstream of the conserved poly(A) site. The region immediately adjacent to the *S. sciureus* core CoTC homologue is not present in humans. The identified stem structure of the core CoTC also seems to be highly conserved: replacement of C31 by U in *Pongo pygmaeus*, and of A169 by G in *S. sciureus*, retains nucleoside pairing important for the structure. Recently, an independent study of the transcriptional termination model (A.R., unpublished data) has been conducted on the basis of a probabilistic generalized linear model and Bayes theorem<sup>17</sup>. The model identifies, through the primary sequence and secondary structure, an additional general termination marker downstream of the poly(A) site in several tested genes. When applied to the human  $\beta$ -globin gene, it successfully identifies the CoTC element as the termination marker of interest. Within the limitations of our current analysis these results encourage us to predict that the autocatalytic cleavage elements within  $\beta$ -globin mRNA might exist as a more general phenomenon. Intriguingly, there exists a biological precedent for a primary ribozyme cleavage downstream of poly(A) sites. In the slime moulds *D. iridis* and *P. polycephalum*<sup>18,19</sup> a complex twin-ribozyme class of group I introns produces a processive autocatalytic cleavage downstream of the poly(A) site of the intron-encoded mRNA. The site of the cleavage does not incorporate GTP, generates RNA products with 3'-hydroxy and 5'-phosphate ends, and is required for mRNA processing and poly(A) site recognition, in a similar manner to the CoTC element.

Earlier results identified CoTC as an essential part of transcriptional termination for the human  $\beta$ -globin gene *in vivo*<sup>3</sup>. We show here that this CoTC activity involves autocatalysis. Cleavage associated with recognition of the poly(A) signal does not promote termination in the absence of CoTC<sup>3</sup>. Previous studies found that poly(A) site cleavage was sufficient to induce termination only when a downstream pause site was present<sup>20</sup>. This suggests that poly(A) site cleavage, unlike CoTC, is kinetically slow. CoTC, but not for example a hammerhead ribozyme<sup>21</sup>, generates 5'-phosphate and 3'-hydroxy ends at the site of cleavage. These are preferred substrates for Xrn2 and the 3'  $\rightarrow$  5' exoRNases within the exosome complex<sup>22–24</sup>. Indeed, the AU-rich stem of the CoTC element, essential for autocatalytic cleavage, might be related to the AU-rich element (ARE)<sup>25</sup> because both RNAs can efficiently and specifically recruit the exosome in a cell-free ATP-dependent RNA decay system (A.A., unpublished data). At the same time, the 3' RNA cleavage product generated by CoTC activity is subjected to 5'  $\rightarrow$  3' exonuclease degradation by Xrn2, which in turn promotes transcriptional termination<sup>26</sup>.

We suggest that regulated autocatalytic cleavage elements might have a more general function in RNA metabolism. Judiciously positioned, these elements might provide specific entry sites for distinct exoRNases, with direct implications for transcriptional termination, intron degradation and mRNA turnover<sup>22</sup>. □

## Methods

### Constructs

All DNA constructs were made by polymerase chain reaction (PCR) amplification of the relevant DNA by *Pfu* Turbo polymerase (Stratagene). Sequence coordinates of the CoTC element refer to the region 64,568–65,370 (GenBank accession no. U01317). For p $\beta$ 8<sub>191</sub>,

the minimal 191 base pairs of the ribozyme were amplified by PCR with oligos B85' and R<sub>200</sub>. This was cloned blunt-ended into a vector prepared by PCR amplification of  $\beta\Delta 5$ –7<sup>3</sup> with oligos VR and VF (this vector will be described as  $\beta\Delta T$ ). For p $\beta$ 38<sub>191</sub> an insert was generated with oligos B85' and R<sub>200</sub>Clal containing a *Clal* restriction site at the 3' end. This insert was cloned blunt-ended into a vector prepared by PCR amplification of p $\beta$ 8<sub>191</sub> with oligos R<sub>200</sub> and VF2. The resulting construct was subject to digestion with *Clal* followed by insertion of another copy of the B85'/R<sub>200</sub>Clal-generated insert. pSKO1 was made by the insertion of a SKO5'/SKO3'-generated PCR product (coordinates 64,634–64,723) into  $\beta\Delta T$ . pSKO2 was made by inserting a SKO5'/SKO3' PCR product into a vector prepared by PCR amplification of pSKO1 with oligos SKO3' and VF2. The  $\beta\Delta 5$ –10 plasmid<sup>3</sup> and Tat plasmid<sup>27</sup> have been described previously. The CoTC element was cloned from the construct<sup>3</sup> instead of the original sequence between the G-less cassettes. Templates for the CoTC RNA and its mutants were generated by PCR with T7 promoter at the 5' end of the upstream primer. Point mutants were made with the QuikChange site-directed mutagenesis kit (Stratagene).

### Transcription *in vitro*

Reconstituted transcription was performed as described<sup>4</sup> in the presence of 5  $\mu$ g of nuclear extracts. After transcription and optional treatment with RNase T1 (10 U for 15 min), transcripts were extracted with phenol/chloroform, precipitated, and resolved on 5% denaturing acrylamide gels. HeLa nuclear extract was purchased from the C4 Cell Culture Center (Belgium).

### RNA production and purification

CoTC RNA was transcribed from the PCR templates with T7 polymerase in the presence of lower concentrations of Mg<sup>2+</sup> and NTP to decrease autocleavage. Transcripts were gel-purified, phenol-extracted and precipitated. Tryptic digestion was performed before phenol extraction as indicated.

### Cleavage

The reactions were conducted at 37 °C in the presence of Tris, Mg<sup>2+</sup>, KCl or NH<sub>4</sub>SO<sub>4</sub> and GTP as indicated. The reactions were stopped with EDTA and 80% formamide and the products of cleavage were resolved on 8 M urea gels.

### S1, T1, V1 and chemical modification analyses

All reactions were conducted as described in ref. 28.

### Transfection and run-ons

Subconfluent HeLa cells were transiently transfected with 20  $\mu$ g of test plasmid and 3  $\mu$ g of the Tat plasmid with the use of Lipofectamine 2000 (Invitrogen) in accordance with the manufacturer's guidelines. Probes  $\beta 3$  and  $\beta 4$  as well as P and U3 (ref. 29) have been described previously. The A singlestranded DNA NRO probe was prepared by the insertion of an *Aat*II/*Xmn*I digestion of  $\beta\Delta 5$ –7 into M13mp19 digested with *Hinc*II. NRO analysis was performed as described previously<sup>30</sup>.

### Bioinformatics

The folding algorithm<sup>13</sup> was modified to incorporate data from the enzymatic and chemical modification analyses. Sequence alignments were performed on the NIH server ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)).

Received 28 June; accepted 17 September 2004; doi:10.1038/nature03032.

- Greger, I. H., Aranda, A. & Proudfoot, N. J. Balancing transcriptional interference and initiation on the *GAL7* promoter of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 8415–8420 (2000).
- Krishnamurthy, S., He, X., Reyes-Reyes, M., Moore, C. & Hampsey, M. Ssu72 is an RNA polymerase II CTD phosphatase. *Mol. Cell* **14**, 387–394 (2004).
- Dye, M. J. & Proudfoot, N. J. Multiple transcript cleavage precedes polymerase release in transcription by RNA polymerase II. *Cell* **105**, 669–681 (2001).
- Kwek, K. Y. et al. U1 snRNA associates with TFIIF and regulates transcriptional initiation. *Nature Struct. Biol.* **9**, 800–805 (2002).
- Lee, M. J. & Greenleaf, L. A. Modulation of RNA polymerase II elongation efficiency by C-terminal heptapeptide repeat domain kinase I. *J. Biol. Chem.* **272**, 10990–10993 (1997).
- Lilley, D. M. J. The origins of RNA catalysis in ribozymes. *Trends Biochem. Sci.* **28**, 495–501 (2003).
- Cech, T. R., Zaug, A. J. & Grabowski, P. J. *In vitro* splicing of the ribosomal RNA precursor of *Tetrahymena*: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487–496 (1981).
- Doherty, E. & Doudna, J. Ribozyme structures and mechanisms. *Annu. Rev. Biochem.* **69**, 597–615 (2000).
- Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–286 (2004).
- Li, Y. & Breaker, R. R. Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2'-hydroxyl group. *J. Am. Chem. Soc.* **121**, 5364–5372 (1999).
- Emilsson, G. M., Nakamura, S., Roth, A. & Breaker, R. R. Ribozyme speed limits. *RNA* **9**, 907–918 (2003).
- Zuker, M. Mfold web server for nucleic acid folding and hybridisation prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
- Winkler, W. C. & Breaker, R. R. Genetic control by metabolite-binding riboswitches. *ChemBiochem* **4**, 1024–1032 (2003).
- Cech, T. R. The chemistry of self-splicing RNA and RNA enzymes. *Science* **236**, 1532–1539 (1987).
- Bass, B. L. & Cech, T. R. Specific interaction between the self-splicing RNA of *Tetrahymena* and its guanosine substrate: implications for biological catalysis by RNA. *Nature* **308**, 820–826 (1984).
- Stage, K. T. & Uhlenbeck, O. C. Hammerhead ribozyme kinetics. *RNA* **4**, 875–889 (1998).



17. Down, T. A. & Hubbard, J. P. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**, 458–461 (2002).
18. Vader, A., Nielsen, H. & Johansen, S. *In vivo* expression of the nucleolar group I intron-encoded I-dirl homing endonuclease involves the removal of a spliceosomal intron. *EMBO J.* **18**, 1003–1013 (1999).
19. Rocheleau, G. A. & Woodson, S. A. Requirements for self-splicing of a group I intron from *Physarum polycephalum*. *Nucleic Acids Res.* **22**, 4315–4320 (1994).
20. Yonaha, M. & Proudfoot, N. J. Transcriptional termination and coupled polyadenylation *in vitro*. *EMBO J.* **19**, 3770–3777 (2000).
21. Samarsky, D. A. *et al.* A small nucleolar RNA: ribozyme hybrid cleaves a nucleolar RNA target *in vivo* with near-perfect efficiency. *Proc. Natl Acad. Sci. USA* **96**, 6609–6614 (1999).
22. Parker, R. & Song, H. The enzymes and control of eukaryotic mRNA turnover. *Nature Struct. Mol. Biol.* **11**, 121–127 (2004).
23. Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102**, 765–775 (2000).
24. van Hoof, V. A. & Parker, R. Messenger RNA degradation: beginning at the end. *Curr. Biol.* **12**, 285–287 (2002).
25. Chen, C. *et al.* AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell* **107**, 451–464 (2001).
26. West, S., Gromak, N. & Proudfoot, N. J. Human 5' → 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* doi:10.1038/nature03035 (this issue).
27. Adams, S. E. *et al.* Synthesis of a gene for the HIV transactivator protein Tat by a novel single-stranded approach using gap repair. *Nucleic Acids Res.* **15**, 4287–4287 (1988).
28. Tahiri-Alaoui, A. *et al.* High affinity nucleic acid aptamers for streptavidin incorporated into bi-specific capture ligands. *Nucleic Acids Res.* **30**, 1–9 (2002).

29. Dye, M. J. & Proudfoot, N. J. Terminal exon definition occurs co-transcriptionally and promotes termination of RNA polymerase II. *Mol. Cell* **3**, 371–378 (1999).
30. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human  $\beta$ -globin locus. *Genes Dev.* **11**, 2494–2509 (1997).

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank T. Nilsen, J. Manley, S. Valadkhan, C. Smith and P. C. Branco for their critical comments and constructive support. This work was supported by grants to W.J. from BBSRC and Edward P. Abraham Research Fund, to N.J.P. from the Wellcome Trust and to A.A. from the Wellcome Trust Career Development Programme, Medical Research Council, Cancer Research UK, Edward P. Abraham Research Fund and Exeter College (Oxford).

**Authors' contributions** A.T. is the lead author. A.T.-A. and S.W. are second authors, and contributed equally to this work. A.T.-A. and W.J. are responsible for the RNA secondary structure analysis. S.W. and N.J.P. are responsible for making the CoTC mutants and testing them by nuclear run-on analysis.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to A.A. (alexandre.akoulitchev@path.ox.ac.uk).

## Contacts

**Publisher:** Ben Crowe  
**Editor:** Paul Smaglik  
**Marketing Manager:** David Bowen

### European Head Office, London

The Macmillan Building  
4 Crinan Street  
London N1 9XW, UK  
Tel +44 (0) 20 7843 4961  
Fax +44 (0) 20 7843 4996  
e-mail: [naturejobs@nature.com](mailto:naturejobs@nature.com)

**Naturejobs Sales Director:**  
Nevin Bayoumi (4978)

### UK/ RoW/ Ireland:

Matt Powell (4953)  
Andy Douglas (4975)  
Frank Phelan (4944)

### Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

**Natureevents:** Silke Opstrup (4994)

### France/ Switzerland:

Amelie Pequignot (4974)

### Production Manager:

Billie Franklin  
To send materials use London  
address above.  
Tel +44 (0) 20 7843 4814  
Fax +44 (0) 20 7843 4996  
e-mail: [naturejobs@nature.com](mailto:naturejobs@nature.com)

### Naturejobs web development:

Tom Hancock

### Naturejobs online production:

Stefan Hales

### European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen  
Tel + 49 89 54 90 57 11/-2  
Fax + 49 89 54 90 57 20  
e-mail: [p.phelan@nature.com](mailto:p.phelan@nature.com)  
[o.wulffen@nature.com](mailto:o.wulffen@nature.com)

### US Head Office, New York

345 Park Avenue South,  
10th Floor, New York, NY 10010-1707  
Tel +1 800 989 7718  
Fax +1 800 989 7103  
e-mail: [naturejobs@natureny.com](mailto:naturejobs@natureny.com)

### US Sales Manager:

Peter Bless

### Japan Head Office, Tokyo

MG Ichigaya Building (5F),  
19-1 Haraikatamachi,  
Shinjuku-ku,  
Tokyo 162-0841  
Tel +81 3 3267 8751  
Fax +81 3 3267 8746

### Asia-Pacific Sales Director:

Rinoko Asami  
e-mail: [rasami@naturejp.com](mailto:rasami@naturejp.com)

# naturejobs

## Out of fashion

David Bowie once described fashion in an eponymously named song as full of “tension and fear”. He was describing clothing, music and dancing, but he might as well have been singing about science.

Most scientists would scoff at such a suggestion, saying that they have no interest in keeping up with the cool kids at the bench next door. This may be true of superficial trappings — although some do have a need to sport the latest flat-screen display or high-throughput technology. But the real trend-setting runs a little deeper — it's the need to keep up with the pack in research topics. And pressure from external forces to make a researcher's science more relevant to society only exacerbates the issue.

It may seem like the best approach, but following trends can be detrimental to your career. A recent article in *EMBO Reports* (K. Weigmann *EMBO Rep.* 5, 1028–1031; 2004) presents a few case studies. After a Nobel prize and the scare over mad-cow disease, prions became ‘flavour of the month’ — which was good news for the few scientists already working in the field. But the huge boost in publicity prompted many others to jump on the bandwagon — and then the funds dried up.

Even if money had remained available, the hype would have attracted more competition, so ultimately everyone would have got less. Even in well-funded fields, it can be dangerous to jump on someone else's train — you might get scooped. But there is a solution. Ask a novel question and either invent a new technique or use the latest methods to answer it, says Bruce Alberts, president of the US National Academy of Sciences (see *Nature* 431, 1041; 2004).

It's much more satisfying to start a trend than follow one. And doing so means you won't end up a fashion victim.

**Paul Smaglik**  
*Naturejobs* editor



## Contents

### CAREERS & RECRUITMENT

Biomedical engineering  
comes of age

p532

### WWW.NATUREJOBS.COM

Career centre  
Information on the  
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

# Material gains

**The marriage of engineering, medicine and biology is offering people from a wide range of disciplines the chance to accelerate their careers. Myrna Watanabe investigates a growth industry.**

**T**he emergence of new interdisciplinary fields is often signalled by prefixes and suffixes such as 'nano' or 'omics'. The application of engineering and materials science to medicine is no exception — here the prefix is just three letters: 'bio'. "A lot of chemical-engineering and chemical departments now put 'bio' in their names," says Robert Langer, a chemical engineer at the Massachusetts Institute of Technology (MIT) in Cambridge. And this rise in popularity for research into biomedical engineering is translating itself into job opportunities. The National Science Foundation (NSF), for example, predicts that in the United States, this sector will expand by 2,000 posts by 2010, up from 7,000 in 2000.

One of the most attractive aspects of this burgeoning field is its diversity. Career opportunities exist for physicists, chemists and biologists just as much as for engineers. Some biomedical engineers, for example, work with both naturally occurring and synthetic polymers in a wide range of projects — creating materials to repair bones or teeth, or designing polymers to improve the 'mouth feel' of milkshakes. Others work at the molecular scale, where applications

stretch from drug delivery to cosmetics and food. Another group specializes in diagnostics such as gene and protein chips, and a sizeable group works in tissue engineering, providing scaffolding for tissue growth or culturing stem cells in bids to generate replacement organs.

The rise of this sector in the United States has been influenced by two main organizations. Over the past 27 years, the Whitaker Foundation in Arlington, Virginia, has pumped about \$615 million into the various disciplines that have been brought together to grow the field. Its goal was to create new academic departments and raise the field's profile so that it could become self-sustaining (see *Nature* **403**, 463–465; 2000). This initiative was built on in 2000 when the National Institute of Biomedical Imaging and Bioengineering was established. Grants from this institute are now funding the departments seeded by the Whitaker, as well as starting others (see *Nature* **425**, 324–325; 2003).

Although a few academic centres exist in Europe and Asia, much of the rest of the world has scrambled to catch up. One of the biggest efforts is Singapore's Institute of Bioengineering and Nanotechnology, which moved to a US\$300-million facility in May.

## Getting the job done

Alison McGuigan (right) holds a UK passport, but when she came to look for a postdoc position she found herself in a difficult situation. Because biomedical engineering is such a hot field, she had a better than average chance of getting a key placement. And as she had trained in one of the field's premier labs — the Institute of Biomaterials and Biomedical Engineering at the University of Toronto — her prospects were even better. But she had spent too much time in Canada to apply for postdocs normally available to Europeans, so she opted for a "fairly aggressive approach" to finding a post.

McGuigan began her search for a postdoc about a year before defending her thesis. She asked her adviser, Michael Sefton, to assist her. She picked the labs she was interested in and targeted people she wanted to work with, and invited them to her presentations at conferences.

Before she completed her application to her first choice, George Whitesides' lab at Harvard University, she set up a website that detailed all of her research experience — information that would not fit on her CV. She put up a teaching dossier that included her teaching strategies and experience. She also made sure that she communicated to the lab the skill sets she already had, what she wanted to learn in a postdoc, and why her skills would be beneficial to the lab. Sefton forwarded her material to Whitesides' lab. McGuigan was accepted as a postdoc about five hours after she applied. **M.W.**



## INDUSTRIAL EVOLUTION

But much of the job growth is in industry and, with a few key exceptions, those jobs are growing faster in the United States and Canada than in Europe, says Eric Abel, head of the biomedical engineering department at the University of Dundee, UK. He notes that the job market in Britain would not sustain an undergraduate department in the field, and some of the large orthopaedic companies have downsized. But, he explains, there is venture capital available for start-ups. In Scotland, for example, he says that the economic development agency Scottish Enterprise is putting money into commercializing products.

With the focus firmly on industry, it is relatively easy to sidestep the traditional tenure track. In the United States, for example, new graduates are finding work at every degree stage, says Kathryn Uhrich, a biomedical engineer at Rutgers University in Piscataway, New Jersey. PhDs are even able to skip postdocs — something that is hard to do in more traditional life-science fields (see *Nature* **432**, 254–255; 2004) — especially if you are willing to work in industry, agrees Hari Reddi, director of the Center for Tissue Regeneration and Repair at the University of California, Davis.

Lauren Ciccarelli, who earned a bachelor's degree in biomedical engineering this spring from Drexel University in Philadelphia, soon found full-time work





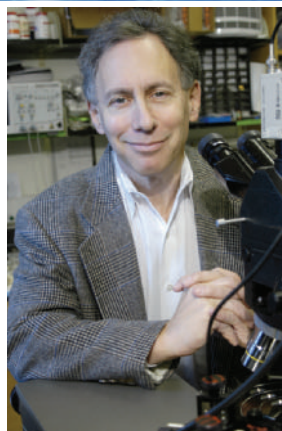
K. GULDBRANDSEN/SPL as an associate engineer specializing in biomechanics at the Philadelphia arm of engineering-consulting firm Exponent. She began working for the company as part of Drexel's undergraduate cooperative education programme. The working relationship proved successful, so Ciccarelli stayed with the firm through two other cooperative programmes and worked part-time during the academic year. Even though she now works for Exponent full time, Ciccarelli is also pursuing a master's degree at Drexel.

Although postgraduates can move straight into work, Uhrich says that most of her undergraduate students go on through academia, pursuing degrees in chemistry or biochemistry, or attending medical school. Engineers with master's degrees and engineering PhDs often go straight into jobs without doing a postdoc, explains Reddi. Someone with an MD degree won't begin a career until he or she is 31 or 32 years old, after finishing a residency and specialized training, he says, whereas in materials science, careers can begin at the age of 23.

#### ACADEMIC APPROACH

Industry might be the most obvious route to take, but it is not the only option for the budding biomedical engineer. There are many open faculty positions, mostly in the United States, says Michael Sefton, director of the Institute of Biomaterials and Biomedical Engineering at the University of Toronto, Canada, and the same top candidates are all chasing them. "The top institutions are all interviewing the same dozen people, and there may be 150 openings across North America," he says.

For those considering doctorates, academic pedigree may be very important. A handful of key labs currently generate the bulk of the available talent. Among those frequently cited by researchers are the labs run by Langer and Reddi, as well as those at the University of Toronto, Johns Hopkins University in Baltimore, Maryland, and the University of Utah in Salt Lake City.



Bioengineered solutions such as this ceramic bone are providing the foundation for an expanding field that Robert Langer (left) feels has huge potential.

To some extent, politics will dictate where biomedical-engineering jobs spring up in the future. Shuguang Zhang, associate director of MIT's Center for Biomedical Engineering, says that Japan shows a lot of promise. Organ transplantation has been allowed in Japan only since 1997 and remains a relatively uncommon practice. As a result, there is tremendous interest in tissue engineering as an alternative, and

several biomedical-engineering programmes have emerged in the country.

China, too, has chosen biomedical engineering as a key area to emphasize as the country builds up its scientific infrastructure. Japan, China and Singapore may have advantages over the United States in that they have fewer restrictions on the use of human embryonic stem cells. This is key to the field of organ regeneration, because these cells could theoretically be used to grow any tissue type and could be manipulated to minimize rejection of transplants by the immune system.

Sefton says that global stem-cell politics and visa restrictions could well play a role in where the jobs of the future arise — and where recruits are willing to go. Many Canadian students have told him that they have mixed feelings about working in the United States — especially after the re-election of President George W. Bush, who has been restrictive in terms of stem-cell and immigration policies.

Although such political considerations may seem to hamper some public-sector work in the United States, the private sector has already shown signs of being resilient. And eliminating some academic opportunities in the United States is likely to create more chances elsewhere in a field that is showing good signs of continued growth. ■

Myrna Watanabe is a freelance science writer in Patterson, New York.

#### Web links

Biomedical Engineering Network  
 ♦ [www.bmenet.org/BMEnet](http://www.bmenet.org/BMEnet)  
 Biomedical Engineering Society  
 ♦ [www.bmes.org](http://www.bmes.org)  
 IEEE Engineering in Medicine and Biology Society  
 ♦ [www.eng.unsw.edu.au/embs/index.html](http://www.eng.unsw.edu.au/embs/index.html)  
 The Whitaker Foundation  
 ♦ [www.whitaker.org/index.html](http://www.whitaker.org/index.html)